An Analysis of the Mobile App Review Landscape: Trends and Implications

Leonard Hoon, Rajesh Vasa, Jean-Guy Schneider & John Grundy
Faculty of Information and Communication Technologies
Swinburne University of Technology
Melbourne, Australia
{lhoon, rvasa, jschneider, jgrundy}@swin.edu.au

Abstract

Context: Apple Inc.'s App Store offers a distribution mechanism for apps and a public review system which allows users to express opinions regarding purchased apps. The ratings and reviews left by users have the potential to influence new users and, hence, have an impact on the commercial feasibility of an app.

Objective: Current literature has extensively investigated reviews of books, movies and hotels. However, there is a limited understanding of reviews for mobile apps. In this work, we analyse a large sample of reviews for top rated apps in order to determine the nature of the reviews, and how these reviews evolve over time.

Method: We performed a statistical analysis of approximately 8 million app reviews to identify the general distribution of review size, the rate of growth for reviews, and analyse the change of both rating and review size over time using the Gini coefficient.

Results: We found that (i) most reviews are short and the majority of apps receive well under 50 reviews in their first year, (ii) apps receive a higher number of short reviews in comparison to long reviews as they age, (iii) around half of the apps decrease in user perceived quality over time, as reflected by the star ratings, and (iv) the rate of review growth and profile of reviews changes significantly between various apps as well as categories.

Conclusion: Developers can use the data presented in this work to benchmark their app. Our approach offers insights for both newer and older apps as we analyse reviews over time. We recommend that developers regularly monitor early reviews and continuously refine their app in order to ensure they meet changing user expectations.

1. Introduction

The App Store for iOS mobile apps allows users to express their level of satisfaction regarding the apps they have purchased through a public review system. Such systems are not uncommon in other domains (*e.g.* online book

Table 1: Examples of Detailed and *Noisy* reviews on the App Store.

Sample Detailed Review	
Essential app for brainstorming ⋆⋆⋆⋆	
In general, the UI does not get in the way, and generating ideas is fast and simple. With the new Apple TV launch and being able to display the iPad on a meeting room TV, I see a lot of potential for team brainstorming. A couple of UI changes would make it a 5 in my opinion: 1) To bring up the delete menu, it's a gimmicky "shake your device". What about a simple delete button on the menu bar, for us uncoordinated senior managers? There's plenty of room. 2) Drawing links between the notes is handy, but it really needs a link feature like iThoughtHD where I can rearrange the notes and the links move with them. 3) The ability to define "note templates". So I could prepopulate a green sticky template with the heading "Opportunity:" and a red sticky template with "Risk:"	
Sample Succinct Reviews	
Great app ⋆⋆⋆⋆⋆	**Drawing fun.** ⋆⋆⋆⋆⋆
Really good app. Very useful	Fun application
Magic piano ⋆⋆⋆⋆	**Good game!** ⋆⋆⋆⋆
I really enjoy this app.	Easy and fun to play!!!

stores, film and hotel reviews). However, this level of access to crowd-sourced opinion is relatively new to the software distribution model and its impact is not yet fully understood. Studies toward understanding online review impact in book, film and hotel domains have resulted in better business intelligence [1, 2] and positive effects on sales [3, 4, 5, 6]. Thus, we expect that the public review system of software distribution channels will similarly influence the purchasing decisions of potential users. Regardless of domain, polarised (positive or negative) reviews *en masse* could reverberate amongst potential users and propagate the general review sentiment exponentially due to the social growth dynamics of crowd-sourced opinions [7].

Apps in the mobile domain generally possess low unit pricing, requiring a high volume of sales in order to have a financially viable product. In this environment, delivery of high quality products from early releases give developers a competitive opportunity for commercial success. Apps also tend to undergo rapid and short iterations of development that add features, correct issues and distribute releases/updates (post a mandatory review process from Apple). The past few decades of software engineering has continuously shown the value of end-user input, and software distribution platforms like Apple's App Store offer a simple method to users to provide this feedback. Given the low friction point, users do express their opinion in reviews, and rate apps. But, what exactly do they say? Since, it is easy to leave a review, do they leave valuable feedback, or just short low-information value reviews? How do these reviews change over time? An empirical analysis of a large sample of reviews can help us answer these questions and allow developers a framework from which they can assess the performance of their apps.

Our cursory manual inspection of App Store reviews shows that users pro-

vide valuable feedback to developers, most commonly by expressing desire for functional extension or warning other potential users of defects. Unfortunately, there are also succinct reviews which, in volume, may be categorised as *noise* (see Table 1 for a sample of detailed and succinct reviews). These short reviews should probably be treated with care in any comprehensive analysis. However, given the relatively immature state of this domain, neither we nor app developers possess a ubiquitous picture of what exactly users like and/or dislike. Neither do we know the general properties of the mobile app review landscape. Specifically, is there useful information, and if so, what proportion of reviews offer useful information? How can we obtain this information while filtering out review noise? Beyond a basic insight into the general types of reviews, we possess neither knowledge of how these reviews evolve over time, nor do we have a model for how the language presented in the reviews relates to the overall rating given. Insight of how this language evolves and, ultimately, how the overall rating evolves would support important future work in this domain.

Existing research efforts in this domain have analysed reviews (sample sizes ranging from 144 to 277,345 reviews) using manual [8] and automated mechanisms involving statistical analysis [9, 10, 11], and supervised machine learning [12] in order to infer information from reviews about user sentiment, product acceptance and reviewer behaviour, respectively. These investigations found that (i) concrete functional and non-functional requirements were derivable from reviews, (ii) a significantly higher range of words were employed in expressing negative opinions than with positive sentiments, (iii) no correlation between price and rating, nor between price and download count was observed, (iv) review frequency and continuity are positively related to rating, comments and helpful votes, and (v) mobile reviews also comprise instances of colloquialisms, intentional misspellings and sarcasm. Although they offer us a broad overview, these studies analysed the data captured on a set date and have not directly investigated review growth and evolution over large time periods. Furthermore, most of these studies investigated small data sets and there still is an important gap in the literature on what the general landscape looks like in terms of how many reviews an app can expect, what quality they are, how to tell quality review from noise, how reviews and ratings change over time, and if this differs significantly across app domain categories.

For our work, we start with the basic question whether user reviews give sufficient information to allow the extraction of deep insights. Also, what is the nature of these reviews? To answer these questions, we analysed 8.7 million reviews of approximately 17 thousand top ranked iOS apps. These apps and reviews were mined across all 22 categories of both Paid and Free price points. More specifically, in this work, we address the following specific questions:

- What is the common size of a user review?
- Does the app rating affect the length of a review?
- How many reviews does an app typically receive?
- How do the number of reviews grow over time?
- Does the category of an app influence the length of a review, the number of reviews, the quality of reviews or review growth over time?

Answering these questions will lay the foundation for developing an app review model and a benchmarking framework that can support market entry testing. Our work is motivated by the hypothesis that techniques that can mine user opinions and their evolution allow developers to prioritise and focus their efforts towards meeting user expectations, and provide decision support for developers looking into developing for specific categories and price points. For instance, if a developer has just launched a new health app, it would be ideal to know how many reviews an app in this category is expected to receive in the first week, and when developers should regularly monitor the app store. Our work and the answers to the research questions that we posed earlier can help app developers better understand the overall landscape that they are competing in. Specifically, a statistical model of reviews and how these reviews evolve over time will offer us a benchmark from which we can compare other apps. Additionally, such a model can also be used to inform future work by better targeting the extraction of samples to help undertake qualitative work to understand sentiment and specific aspects that users like and dislike in apps.

The rest of this work is structured as follows: In Section 2 we explore relevant work in this area. Following that, we outline the experimental setup of our study as well as the specific techniques we used to analyse app reviews in Section 3. The results of our approach are then presented in Section 4, preceding our discussion with respect to our research goals (Section 5). Finally we draw conclusions from our work to date and propose future directions in Section 6.

2. Related Work

User reviews behave like online word-of-mouth (WOM), which is recognised as influential in information transmission, particularly with experience goods [13, 14]. Consumer or user generated online reviews implicitly communicate user-perceived quality based on actual usage experience and satisfaction [15], from which perceivable ease of use, usefulness and ultimately, acceptance [16] are inferable. This creates a feedback loop, providing an opportunity for insights as presented in [8] towards refinement in subsequent iterations of release.

Broadly speaking, this feedback loop alludes to co-value creation [17]. In that, value can be contributed by users through (i) providing feedback, thereby raising the strengths and weaknesses of an app to both the community and developers; (ii) positive feedback, generating or sustaining user positive awareness of the app, thus improving an app's chance of discovery by other users (which is good for developers), raising potential user confidence in the app especially if there are a large number of good reviews, and can impact on the commercial well-being of the developer by presenting an impression of product quality; and (iii) negative feedback, which serves to publicly expose areas of improvement to the developers as well as highlight issues to potential future users.

In effect, the information value of reviews depends on perspective. Positive reviews are valuable to and can be enjoyed by developers, while negative reviews add more value to potential future users. Although negative reviews may be calamitous for developers, it informs prioritisation of their efforts. The resulting decrease in downloads from bad reviews does however, limit the app's exposure, reducing the potential negativity of future reviews.

Despite the concept of reviews for mobile apps being relatively new, other industries have incorporated user feedback into their business strategy for a number of years. The structure of reviews is similar across domains, where users typically can leave an ordinal rating (often as stars) to express sentiment, and a brief text or audio/video comment for justification. Additionally, quality control mechanisms such as "helpfulness" voting may be employed, which was examined by Kostakos in [18], showing that these mechanisms can be useful for prioritising crowd-sourced reviews that other users have rated as helpful for presentation. They also have the potential to polarise the reviews (either too positive or too negative) if users are forced to leave a quality review. However, Chen and Huang [11] found this "helpfulness" mechanism was positively associated with review frequency and continuity, and that review size presented a negative correlation with continuity, but enjoyed a positive correlation with review frequency.

User reviews have influenced sales [3] and consumer preferences [1]. Duan, Gu and Whinston [19] captured the relationship between online reviews and sales using a statistical and temporally segmented approach to study movie reviews and found that users tended to only sample reviews. They also observed the significant influence that box office revenue and review volume have on review valence [5]. Within the hotel industry, proper analysis of reviews was found to yield insightful business opportunities [6]. However, as the App Store offers Free and Paid apps, the influence of cost on reviews has to be examined prior to adopting such approaches.

In prior work pertaining to the mobile app review domain, Gebauer, Tang and Baimai [8] identified that user reviews could be manually mapped to software quality attributes and user requirements. It is also acknowledged that reviews often addressed myriad aspects within the context of the domain and the object under review [8, 19], which presents a need for context and domain jargon aware summarisation and spell correction. Plazter, on the other hand, proposed a motive-construct mapping of review text using supervised machine learning in order to derive user motivations [12]. Platzer's work suggests that users will leave reviews of varying length (often domain specific) and that reviews possess many instances of abbreviations, colloquial expressions, irony and non-standard spelling. Chandy and Gu [20] highlighted the presence of spam and sockpuppet reviewers which skew aggregated reviews with malicious intent. Harman, Jia and Zhang [10] employed an automated approach to extract business intelligence comprising of statistical analysis and data clustering to investigate correlations between download count, rating and price on the Blackberry App Store.

Given the disparity of review lengths, approaches from text summarisation of Twitter feeds have also been explored. Sharifi, Hutton and Kalita proposed algorithms to automate micro-blog summarisation [21, 22], while Leskovec, Backstrom and Kleinberg [23] employed temporally aware data clustering and topical extraction approaches to track the rhythms of news cycles. In a similar manner, our work detects generalizable growth patterns using temporal windows as a first step towards employing such an approach for feature extraction.

Analysis of reviews from other domains highlights the significance of review data in volume and the need to account for changes over time. However, to date, no specific studies have sought to understand mobile app reviews with a large data set, nor have to considered how these reviews change over time. Additionally, the non-standard English expressions used in reviews, potential

for spam and *en masse* short reviews of similar sentiment further complicates the information summarisation process when mining raw user reviews. This highlights a need for some form of raw text treatment prior to application of vocabulary analysis techniques. Although we do not seek to undertake a qualitative analysis of reviews in this work, such an analysis can be better informed by understanding the nature of reviews in general and how these reviews change over time. Having a strong model of reviews allows us to better target any text summarisation as well as help pick better samples for any qualitative analysis.

3. Experimental Setup

In this section, we discuss our experimental setup as well as the specific techniques we used to analyse the app review landscape. We start by describing the dataset mined from the App Store in July 2012, followed by a discussion of how we determined the age and size of each review for our analysis. We then describe how our review growth analysis on these mined app reviews was performed. Finally, the techniques used to perform the analysis of the evolution of app ratings and review size are described.

3.1. Dataset

The App Store has enjoyed over 40 billion downloads in less than five years, offering more than iOS 775,000 apps to over 500 million user accounts [24]. For our investigation, we extracted reviews and star ratings from the *top 400* Free and Paid apps in each of the 22 App Store categories listed in Table 2. We obtained these reviews from only the top 400 apps per category as Apple publishes details for these top apps in a format that permits downloading the data without our scripts being flagged as malicious. Hence, our findings are constrained to successful or popular apps, based on their ranking at the time of data collection (last run of the scraper was on 20th July 2012).

Table 2: All app categories available on the Apple App Store. This omits the "All" category, a list of the top apps from all categories.

Apple App Store Categories							
1	Books	7	Games	12	Navigation	18	Social
2	Business	8	Health &	13	News		Networking
3	Catalog		Fitness	14	Newsstand	19	Sports
4	Education	9	Lifestyle	15	Photo & Video	20	Travel
5	Entertainment	10	Medical	16	Productivity	21	Utilities
6	Finance	11	Music	17	Reference	22	Weather

The *Catalog* and *Newsstand* categories were omitted from our study as both categories had less than 40 apps at the time of our data collection and hence offered a small data set for the comparative statistical analysis that we undertake. Though limited to the top apps, our dataset is still fairly large and comprises 8,701,198 reviews left by 5,530,025 users across 17,330 apps on the Apple App Store. Furthermore, since developers will aim to build a successful app (and customers prefer to download these) our findings will offer a valuable insight.

Each app review is primarily comprised of a star rating between 1 and 5, a review title, and a review body. Additionally, we also captured the date that

Figure 1: The cardinality of a Review, left by a User, per Release of an App in a Category.

each review was created, the user that authored the review, and the particular release of the app that it is associated with. The relationships of a review to other entities is illustrated in Figure 1. For our review analysis, in order to capture all of the information provided by users, we appended the body of the review with its title and treated this as the total text of the review.

We compute the age (in days) of a review for each app in our dataset with respect to the very first review that an app has received (based on the complete time-stamp). Hence, the age of a review is always greater than zero and is used in our analysis as a proxy for time.

3.2. Review Size Distribution Analysis

A manual study performed on a small dataset by Gebauer, Tang and Baimai [8] observed that mobile app user reviews tended to be short. Due to the manual approach and small data set size, no strong numerical boundary of review length expectations could be determined. A good understanding of the expected properties of a review can help identify atypical ones, and this awareness can be applied to flag potential spam as well as to track and monitor if user engagement is abnormal. To address the research goal, in this work, we analyse reviews in order to determine the average or typical size of a review, and to check if these sizes are applicable across categories and ratings.

Our metric for size in this experiment is the summed character count of the review title and body. We use the size of the review to approximate user engagement, in that we consider size to resemble the affect that an app has on a user. We analyse the data initially via the use of summary statistics, box plots, and cumulative distribution charts (to gain a visual perspective). We later determine the influence that the rating as well as the category have on the size of the review by using a one-way ANOVA test – where review size is the dependent variable, and rating/category are the independent variable.

3.3. Review Growth Analysis

The questions that we wanted to explore in our study included: How many reviews can an app expect? Is the number of reviews different across categories? Does the growth of review count over time exhibit a generalizable pattern?

In order to answer the first two questions, we determined the median as well as the 75th percentile of reviews within each category, captured at different time intervals. In our work, we consider 1, 6, and 12 month intervals as a starting point to identify the typical growth pattern at a category level. This information offers us a general insight into what an app within a category can expect in terms of reviews. For example, our analysis showed that in the *Business* category, the median value for an app is 33 reviews after a period of 12 months.

Across categories, the median and 75th percentile values provide a broad range of what an app developer can expect from a particular category. However, since this range is a snapshot of a different point in time it does not show how fast reviews grow, nor if there is a generalizable pattern to this growth.

There is no prior literature (to the best of our knowledge) that is available to offer us an insight into the growth rate for reviews. We can broadly assume that growth of reviews would map to the number of downloads over time. Hence, growth may fit a linear trend line if it attracts users at a consistent rate. We would see a sub-linear trend if the app is attracting fewer users over time, while there would be a super-linear growth if the app is a hit and there is an exponential growth in the downloads (within the period of study). In the best case, all apps would fit on of these three growth trends, but realistically we can expect that growth in many instances can be erratic and may happen in spurts with long periods of no growth. In order to better understand the underlying growth dynamics within the data set, we construct regression models with *age* as the independent variable and *review size* as the dependent variable. We construct both linear and quadratic regression models as this combination is needed to determine if the growth is either Linear, Sub-linear (the x^2 parameter in the quadratic model will be negative), or Super-linear (x^2 parameter will be positive) [25].

Given the size of our data set, we anticipated that we would not be able to generate a valid regression model for all apps. Furthermore, we are unable to visually inspect and validate thousands of models manually, and hence we used the following criteria to determine models that can be considered valid: (i) the model had an r-squared value greater than 0.8, (ii) the model satisfied the heteroskedasticity assumption (we applied both Cook-Weisberg test as well as White's test for heteroskedasticity), and (iii) the residuals from the model were normally distributed (Shapiro-Wilk test). The regression models construction as well as the post-estimation tests for heteroskedasticity and residual normality were computed using Stata (p-value$<$0.05).

We applied the above criterion to both the linear as well as the quadratic regression models. If one of the models was rejected in the above step, then the remaining model was considered as the best fit. However, when both models passed all criteria, we used BIC (Baysian Information Criterion) to select the most appropriate model (*i.e.* selecting the model with the lower BIC value as recommended in [26]). In case that neither of the two regression models offered a valid fit, we categorized these apps separately to indicate that a growth model could not be determined (marked in Table 4 as NIL).

Our initial models showed that in most cases apps do not receive reviews every day and hence, we were not able to generate any valid regression models for these apps. In order to ensure that we were able to build useful growth models, we counted reviews at age intervals of *28 days* (approximately a month). The value of 28 ensured that approximately 80% of the apps showed review count growth. Using the longer interval across all apps ensured consistency when evaluating the findings, and also permitted comparisons across categories as well as apps. Additionally, we removed apps that had less than 6 months of evolution from the regression modelling as they did not have sufficient time data to construct an effective model.

3.4. Rating Distribution Evolution

The star rating, however crude and subjective, offers an insight into how users perceive an app. We assume that developers will aim to obtain reviews of the highest possible rating value in order to increase sales and usage of their apps. Hence, we wanted to answer the following two key questions: (i) what are

the review ratings distributions across apps and (ii) do apps tend to improve their ratings over time?

In order to understand how review ratings change over time, we initially aimed to observe the change of *median* rating at regular intervals (we used the 28 day interval in order to be consistent with the analysis undertaken for review size). However, we observed that for the majority of apps, the median rating value does not change sufficiently over time. This stability is not completely surprising given the size of the scale (5 possible star ratings only). Given this limitation of median rating, we observed the *mean* rating at 28 day age intervals to determine if there was a statistically observable monotonic relationship between the mean rating and time. We computed this using the Spearman's correlation coefficient. If apps generally improve their ratings over time, we should observe a positive Spearman coefficient. Conversely, a negative value would indicate that the mean ratings are decreasing over time. To ensure that the dependency identified was sound, we counted apps only when the p-value was less than 0.05, and if the ρ value was greater than 0.5 or less than -0.5. In Section 4, we will report our summary of both when the Spearman correlation was able to determine dependency and where it showed that parameters were independent.

Review ratings fall on a 5-point ordinal scale and hence, the mean is not an ideal summary statistic from a measurement theoretic perspective [27]. However, given the limitation of using medians as discussed above, and since the App Store shows an average value computed from across the ratings, we used a mean value instead as it offers an insight into potential changes over time as communicated to users by Apple.

3.5. Review Size Distribution Evolution

Similar to the distribution of rating values, we wanted to get further insights into how the *size* of reviews change over time. More specifically, we wanted to address the questions (i) how do review size distributions change over time and (ii) do apps tend to receive longer or shorter reviews as they age?

In our early analysis, we observed that review sizes follow a highly skewed distribution and hence, there is a need for a summary statistic that works well with such distributions. We can use the average *word count*, but it does not provide direct insight into the underlying mechanics of how the distribution changes. For instance, we are unable to determine if apps tend to receive shorter reviews as they age without investigating the underlying histogram directly.

In order to answer these questions, we observed how the Gini coefficient [28] of the review size distribution changes over time (again using 28 day intervals). The Gini coefficient is an inequality measure that is commonly used in the social-economics domain to identify if the rich are getting richer and has been shown to be effective when dealing with skewed distributions [29]. Additionally, it is not influenced by the underlying population size making it a good measure to use when comparing different apps and how the wealth distribution changes over time [25]. The Gini coefficient is a value between 0 and 1. Zero implies a perfect equality in the distribution (that is, all reviews in the population have the same word count). A Gini value of 1 implies that the distribution is perfectly inequal (that is, one review has word count, while the rest have a zero word count).

We observed how the Gini coefficient of the review size distribution changes over time (28 day intervals) using a Spearman correlation coefficient. Similar

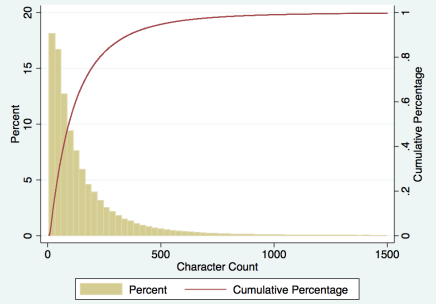

Figure 2: Histogram of Character Count of Reviews (cumulative values are shown on the right hand y-axis)

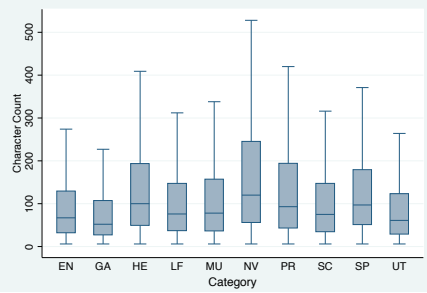

Figure 3: Boxplot depicting Character Counts in relation to Categories. The following categories are depicted. Entertainment(EN), Games(GA), Health & Fitness(HE), Lifestyle(LF), Music(MU), Navigation(NV), Productivity(PR), Social Networking(SC), Sports(SP), and Utilities(UT).

to the method described in the previous section, we set the p-value at 0.05, and selected models only when ρ is either under -0.5 or over 0.5.

As an app ages, if it receives a higher proportion of longer reviews (*i.e.* wealthy reviews from an economic perspective), then the overall population becomes wealthy and this is reflected in a lower Gini coefficient. Conversely, if an app receives a higher proportion of shorter reviews as it ages, then the Gini coefficient will trend higher as it will reflect the greater inequality. If an app receives a similar distribution of reviews over time then the Gini value will remain stable.

The Gini coefficient is population size agnostic. However, the computation involves measuring the area under a curve [28] and hence the values tend to fluctuate significantly when there is a small population size. In our early analysis, we determined that when apps received more than 10 reviews, the Gini coefficient offered a greater level of stability in its movements. Consequently, we observed the change in the Gini value (over time) once an app had received at least 10 reviews. This adjustment implied that in approximately 50% of apps we had to select after they were at least 2 months old (56 days) for their first data point. Furthermore, to ensure that the Spearman correlation coefficient has sufficient data to produce a valid result we removed apps that had less than 6 data points for the time component.

4. Findings

4.1. Typical Size of a Review

What is the typical or average size of a review that users leave? Are users influenced by the category of the app when they leave a review? Does the rating affect the size of the review? We observed that user review length is highly skewed (see Figure 2) with an average of 117 characters, and median at 69 characters (SD: 156, and Skew: 7.24). The data confirms the prior finding from a smaller sample by Gebauer et. al. [8] that users tend to leave short messages – in fact, nearly 75% of the reviews are short enough to fit within a tweet (140 characters).

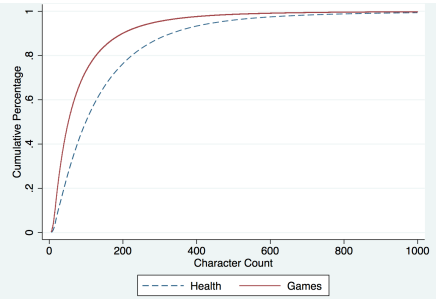

Figure 4: Cumulative percentage of Character Count for the Health & Fitness and Games Categories.

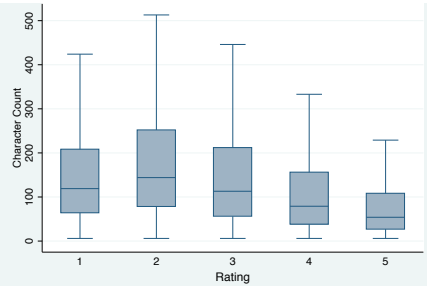

Figure 5: Boxplot depicting Character Counts in relation to Ratings left by users. Outliers (top 2.5% of data) have been supressed to improve readability.

Interestingly, a small set of the reviews (2.5%) are relatively long with over 100 words or 500 characters suggesting that some users do take time to leave long reviews with potentially useful content. This observation shows that review sizes do not present a typical size and the long tail implies that we cannot just rely on a simple summary statistics like the mean. The volume that short reviews occur in suggest that users maybe creating the reviews on a mobile device – given the nature of under input on smartphones it is reasonable to expect a tendency to create short reviews rather than longer ones.

An initial manual inspection did not indicate that users would behave differently across categories (at least with respect to the review length). Given the size of our data set, we expected review sizes to be similar across categories due to the large numbers of apps and reviews. However, the ANOVA test clearly demonstrated that review lengths differ significantly across the 22 categories ($p < 0.01$). This observation is depicted in Figure 3. We depict only 10 randomly selected categories due to space constraints for readability, but the rest of the categories have a similar spread. Interestingly, we found that users tend to leave significantly shorter reviews for *Games* than for other categories.

We also unexpectedly observed a large disparity of the median review size in the *Health & Fitness* category of a 100 characters, which is twice as long as the median for *Games*. This observation is visually summarised in Figure 4 and contrasts the significant difference across the board between the review length of these two categories. A detailed analysis that compared the categories with each other both visually and also by using a Bonferroni comparison (after the ANOVA) showed that in general, most categories presented different review length spreads. A few categories such as *Media* and *Finance* had similar spreads, but in general the category of the app appeared to be influenced by review size.

In order to determine if app rating affects review size left by users, we first performed a manual, cursory scan of reviews on the App Store to find that higher ratings were accompanied by short reviews (often a single word like *awesome*). This is inverse of lower ratings, where we observed discontented users left longer reviews expressing their critique of an app.

We present these observations visually in Figure 5 as a box-plot, depicting that longer messages were left by users when they rated an app poorly (1 or 2 stars). Interestingly, 2 star ratings tend to receive longer reviews than 1 star

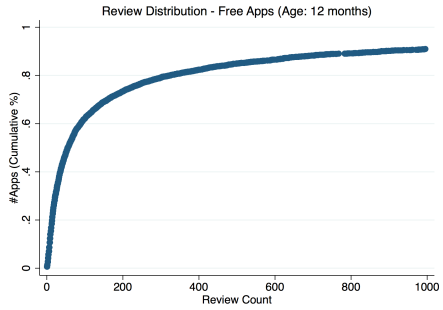

Figure 6: Review counts mapped to the number of apps that receive them for Free apps.

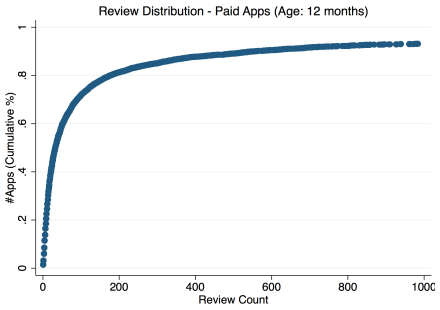

Figure 7: Review counts mapped to the number of apps that receive them for Paid apps.

reviews. We find that users appear to take the time to express their discontent by writing a longer review, in contrast to leaving a succinct review when content. The median size of a 5 star review is 54 characters, while it is nearly three times longer at 144 for a 2 star review. Furthermore, an ANOVA test supports the observation that review lengths differ significantly ($p < 0.01$) across the 5 rating levels.

This indicates that short reviews with a high positive sentiment have the potential to appear in volume, increasing the challenge of information extraction for app strengths due to the lack of content to mine. However, with the higher likelihood of larger sized negative text justification in reviews, we are presented with more raw data to work with in determining app pitfalls to avoid.

4.2. Review Count Expectations

Review counts offer an indication of user engagement, but, how many reviews can an app expect? Do the review count numbers differ significantly across categories?

Our findings for these questions are presented in Figures 6 and 7. We plot the cumulative distribution of apps against the review counts that they receive after 12 months. These plots show that Free apps tend to receive a greater number of reviews compared to Paid apps. For instance, half of the Free apps receive at least 50 reviews, while half of the Paid ones can expect only 30 reviews after 12 months in the App Store. Though we do not have specific numbers on how many users acquire Free and Paid apps, the review count value suggests that users of Paid apps are relatively engaged since we can assume a significantly higher proportion (potentially 4 or 5 times the number) of users install Free apps when compared to Paid apps. Yet, the overall review numbers suggest that even with fewer active absolute users, Paid apps entice a greater level of feedback from users.

A break down by category offers further insight and is summarized in Table 3. This table shows the median and 75th percentile value for the review count at three different age points (1 month, 6 months, and 12 months). This table offers a benchmark on what to expect in terms of the number of reviews. For instance, a Paid *Entertainment* app that has received 40 reviews in the first month can be considered to be performing relatively well, as it falls in the top 25 percentile. However, a Free *Entertainment* app with 40 reviews in its first month is only

Table 3: Review Counts per Category (Free or Paid) at Median and 75th Percentile in 1, 6 and 12 month intervals.

Review Counts per Category												
Benchmark for:	**Free Apps**						**Paid Apps**					
		50%			**75%**			**50%**			**75%**	
Category	**1 mth**	**6 mth**	**12 mth**	**1 mth**	**6 mth**	**12 mth**	**1 mth**	**6 mth**	**12 mth**	**1 mth**	**6 mth**	**12 mth**
Books	4	15	28	9	41	72	4	13	22	10	34	59
Business	5	18.5	33	12.5	39.5	76	5	15	26	11	36.5	74
Education	5	18	40	14	54	94	4	13	24	10	35	60
Entertainment	32	173	302	151	608	739.5	11	54	110	35	202	353
Finance	9	28	40.5	25	74	119	4	12	17	8	31	52
Games	225	585	714	1128	2851	2533	168	603	938	510	1817	2590
Health & Fitness	6	28	52	15	75	158	5	23	43	12	57	122
Lifestyle	13	52	110	34	152	269	5	17	29	13	44	95
Medical	3	9	14	6	21	35	3	10	17	6	21	36.5
Music	12	43	75	43	150	226	8	34	58	22	81	156
Navigation	4	13	23	9	28	53	4	13	22	9	28	53
News	13	63	165.5	41	234.5	624	6	24	47	14	71	160
Photo & Video	8	24	41	21	75	107	3	7	13	6	22	34
Productivity	12	48	96	32	139	384	11	45	88	31	108	205
Reference	9	41	84	26	157	271	8	36	70	22	111	202
Social Networking	7	27	49	19	74	127	5	15	23	10	35	55
Sports	8	27	46	22	68	122	3	9	12	7	20	31
Travel	6	23	39.5	18	51.5	103	3	8	13	6	20	34
Utilities	12	55	102	44	222	344	7	26.5	48	19	83	171
Weather	3	9	16	8	21	33	3	8	12	7	20	31
Total	8	29	52	27	105	191	5	17	30	14	55	104

performing close to the median number of reviews. Though these values are not a direct reflection of the actual users that an app has, it provides an insight into the level of engagement and connection that an app has made with its users.

Certain categories exhibit higher engagement than others, in particular games. We postulate the number of reviews in the *Games* category to be reflective of the proportion of downloads. That is, more users download games and hence they attract a greater absolute number of reviews. This category is also likely to be influenced by factors such as demographics (younger and potentially more engaged audience who are comfortable leaving reviews via the smart phone) as well as active in-app encouragement for reviews, possibly complemented by incentive mechanisms (*e.g.* "Rate us for free in-game currency").

In our previous work [30, 9] we found that Paid apps in general elicit longer reviews, and hence anticipated that these Paid apps may also receive more reviews. Our current data shows that Free apps acquire more reviews in the median and 75th percentile and hold this advantage over time. The simplest explanation is that there are more users of Free apps and hence in absolute terms more reviews. However, in many categories Paid apps receive almost as many reviews as Free apps. This indicates that when users pay, we can expect a greater level of commitment towards providing feedback.

Categories such as *Entertainment*, *Lifestyle* and *News* seem to attract relatively higher review counts during the apps's life cycle. Again, this is likely to be reflective of the underlying number of users. We also postulate that the numbers be amplified by the nature of such apps in that they are designed for content consumption, as opposed to creation of content. In these cases, apps have to differentiate themselves from competitors based on content quality, usability, and other aspects. The consequence of this is a more competitive landscape, raising user expectations for apps in such categories and hence driving more users to leave reviews. This phenomenon may also be at work in the *Games* category.

Certain categories, such as *Medical* and *Weather* apps, exhibit relatively low review counts. Although it appears that this phenomenon may be driven by a

Table 4: Proportion of App Review Growth per Category exhibiting Super-Linear, Linear, Sub-Linear and non-fitting (NIL) trends over Time.

App Review Growth per Category										
Regression Fit for:		**Free Apps**					**Paid Apps**			
Category	***Apps**	**Super %**	**Linear %**	**Sub %**	**NIL %**	***Apps**	**Super %**	**Linear %**	**Sub %**	**NIL %**
Books	228	21.49	24.56	23.25	30.70	253	20.16	20.95	29.64	29.25
Business	256	23.05	26.95	17.97	32.03	272	22.79	24.26	14.71	38.24
Education	253	28.06	24.90	25.69	21.34	265	24.15	27.55	21.89	26.42
Entertainment	236	17.37	18.64	24.58	39.41	247	15.79	22.27	21.05	40.89
Finance	228	17.98	29.82	16.67	35.53	264	16.67	23.48	20.08	39.77
Games	155	7.10	22.58	18.71	51.61	259	5.02	20.85	27.03	47.10
Health & Fitness	263	28.90	22.81	20.15	28.14	269	30.11	22.68	14.50	32.71
Lifestyle	261	22.22	19.54	19.54	38.70	239	19.25	22.59	24.69	33.47
Medical	248	18.95	25.81	20.16	35.08	276	17.03	25.00	23.19	34.78
Music	253	22.13	20.55	24.90	32.41	265	16.23	25.66	24.15	33.96
Navigation	269	25.65	23.05	26.39	24.91	294	21.77	20.75	24.15	33.33
News	252	19.84	21.43	23.81	34.92	250	18.40	24.00	20.00	37.60
Photo & Video	207	21.26	24.64	26.09	28.02	247	24.70	25.10	19.84	30.36
Productivity	257	20.62	24.51	26.46	28.40	261	22.61	22.61	19.92	34.87
Reference	249	24.90	21.29	24.50	29.32	238	18.91	21.85	27.31	31.93
Social Networking	258	24.81	25.19	18.22	31.78	247	17.41	23.48	23.08	36.03
Sports	235	23.40	25.96	16.17	34.47	228	23.68	25.88	18.42	32.02
Travel	238	21.85	24.37	18.49	35.29	314	28.34	22.61	17.20	31.85
Utilities	254	20.47	22.05	21.26	36.22	261	18.77	25.29	25.29	30.65
Weather	201	14.93	33.83	32.34	18.91	236	19.92	28.39	15.68	36.02

* denotes the number of Apps analysed.

lack of users, the *Weather* apps are used widely and hence there would seem to be an additional driver that limits the number of reviews in this category. An explanation for the *Weather* category is that apps in this category can broadly be considered to be presenting information that can be consumed at a glance. Hence, they receive user attention only in very short bursts, perhaps reducing the potential for developing a deeper emotional connection with these apps. This would seem to impact on the number of reviews received. Interestingly, in our earlier work [9] we found that *Health & Fitness* and *Medical* category apps tend to receive comparatively longer reviews. Hence, the comparatively lower overall review count was unexpected. Unlike weather apps, *Medical* and *Health & Fitness* category apps are usually much more complex. Hence, the lower review count is potentially driven by the overall lower number of users downloading apps from these categories.

4.3. Review Growth Trends

In the previous section, we considered how many reviews an app can expect. In this section, we address the question of how the review counts grow and if there is a generalizable pattern, respectively.

In Table 4, we summarize the results from our regression fit analysis with the "*Apps" column indicating the number of apps we analysed. We find that growth does not exhibit a consistent predominant growth pattern. That is, super-linear, sub-linear, linear and non-linear are all possible, and no single trend shows a significantly higher likelihood. Given the variation in growth rates across apps, there is no reliable way to indicate a broad expectation for a given app on what to expect in a category. In effect, this depends predominantly on the app itself and how the developers engage their users.

Before collating our observations, we anticipated that the sub-linear trend may be more predominant, assuming dynamics as expected by Lehman's laws [31] to be at play. However, a closer inspection of the growth data and trend plots showed that different apps exhibit different growth dynamics. Figures 8, 9, and 10 are plots of apps with super-linear, sub-linear, and linear growth in review counts, respectively. Each data point in these figures is comprised of the

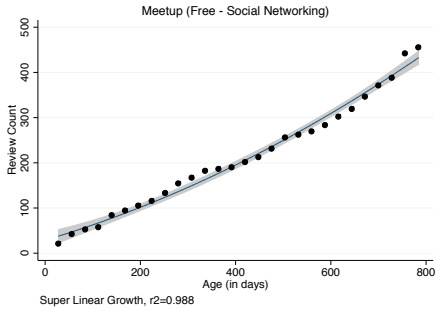

Figure 8: Example of an app exhibiting super-linear growth rate based on review count over the age of the app in days.

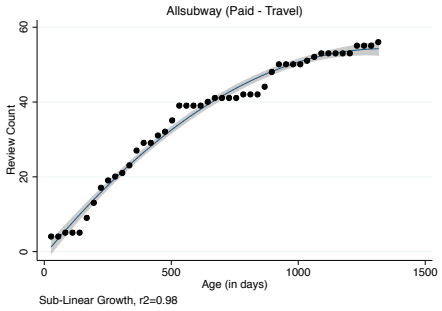

Figure 9: Example of an app exhibiting sub-linear growth rate based on review count over the age of the app in days.

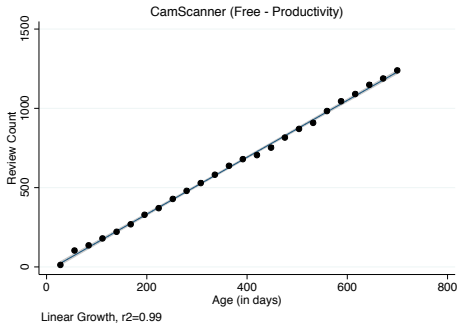

Figure 10: Example of an app exhibiting linear growth rate based on review count over the age of the app in days.

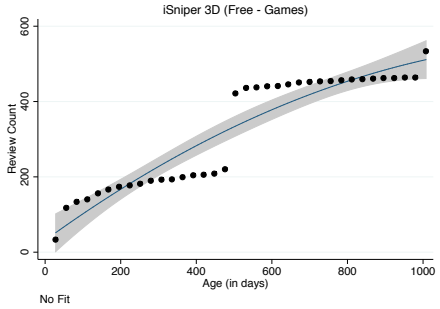

Figure 11: Example of an app that does not exhibit a generalisable growth rate based on review count over the age of the app in days.

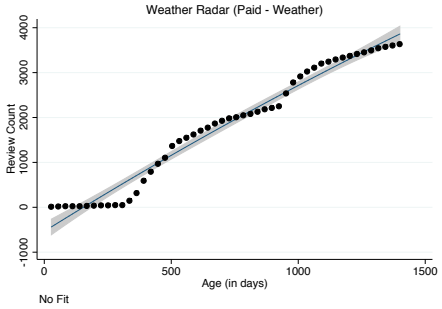

Figure 12: Example of an app that does not exhibit a generalisable growth rate based on review count over the age of the app in days.

review count at each 28 day interval. There are notable spikes or "steps" in review growth that often coincide with new releases of an app, though not always. These steps may have also been driven by marketing campaigns that resulted in growth of user numbers. Interestingly, the longer an app has existed, the more likely it will not exhibit super-linear growth in reviews, suggesting competitive pressures where apps eventually lose popularity.

Approximately 30% of apps in every category fit neither a linear or quadratic

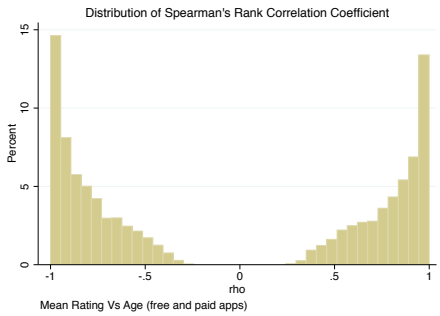

Figure 13: Spearman Histogram of mean rating.

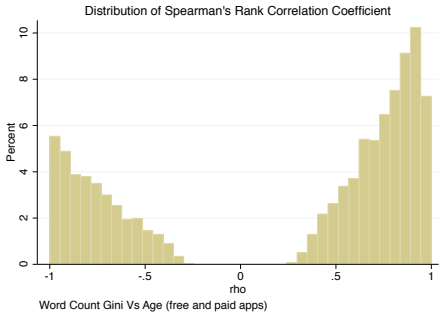

Figure 14: Spearman Histogram of word count gini.

Table 5: Proportion of Apps exhibiting positive and negative ρ values of Mean Ratings and Word Count Gini Coefficient over Time.

Spearman Proportions of Mean Ratings and Word Count Gini				
	Free Apps		**Paid Apps**	
	Mean Rating	**Word Count Gini**	**Mean Rating**	**Word Count Gini**
Category	**Pos. / Neg.**	**Pos. / Neg.**	**Pos. / Neg.**	**Pos. / Neg.**
Books	43% / 57%	63% / 37%	48% / 52%	54% / 46%
Business	37% / 63%	65% / 35%	49% / 51%	66% / 34%
Education	46% / 54%	67% / 33%	42% / 58%	62% / 38%
Entertainment	51% / 49%	57% / 43%	42% / 58%	64% / 36%
Finance	35% / 65%	54% / 46%	46% / 54%	75% / 25%
Games	55% / 45%	67% / 33%	50% / 50%	53% / 47%
Health & Fitness	52% / 48%	64% / 36%	62% / 38%	75% / 25%
Lifestyle	48% / 52%	67% / 33%	48% / 52%	70% / 30%
Medical	58% / 42%	74% / 26%	45% / 55%	62% / 38%
Music	51% / 49%	62% / 38%	45% / 55%	71% / 29%
Navigation	42% / 58%	65% / 35%	46% / 54%	74% / 26%
News	33% / 67%	55% / 45%	29% / 71%	63% / 38%
Photo & Video	53% / 47%	67% / 33%	56% / 44%	65% / 35%
Productivity	54% / 46%	69% / 31%	56% / 44%	76% / 24%
Reference	49% / 51%	59% / 41%	59% / 41%	65% / 35%
Social Networking	48% / 52%	67% / 33%	33% / 67%	62% / 38%
Sports	34% / 66%	56% / 44%	44% / 56%	68% / 32%
Travel	46% / 54%	70% / 30%	54% / 46%	75% / 25%
Utilities	53% / 47%	63% / 38%	57% / 43%	67% / 33%
Weather	43% / 57%	60% / 40%	57% / 43%	67% / 33%

trend (labelled as NIL in Table 4). In Figures 11 and 12 we present examples of such apps. In Figure 11, there is a very clear spike in growth that coincides with a new release, and some segments exhibit an almost sub-linear growth pattern. The same phenomena is also visible in Figure 12, albeit in a much more subtle way. Again, the segmented growth segments coincide with new releases. In both cases, a temporally segmented approach of linear and quadratic regression fit is likely to yield a valid fit on a per segment basis.

These findings suggest that there are a number of apps where users offer feedback resembling a punctuated equilibrium pattern. Hence, developers are advised to monitor user feedback immediately after a release, as that is the time period when they are likely to see a spike in new review comments.

4.4. Review Rating Evolution

The star rating offers insight into how users perceive an app. In this section, we address our next research question – do apps tend to improve their ratings over time?

Our analysis (as explained in Section 3) observed the change in the mean rating over time using the Spearman rank correlation coefficient. Across all categories, we were able to generate valid Spearman correlation coefficients (between mean rating and Age) for 71% (5043) of the apps. We were not able to generate a statistically strong fit for the remaining (29%) due to erratic movements in the mean ratings, occasional spike caused by a new release, as well as general stability in the overall rating over the observed period. In the cases where we were able to generate a statistically useful Spearman coefficient, the value was similar across both Free and Paid apps (see Table 5).

Across these apps with a valid Spearman correlation coefficient, 48% show a positive ρ value. That is, the mean rating increases as the app ages. 52% of apps showed an inverse relationship. With a close to 50% divide of positive and negative relationships, apps are as likely to improve in quality from the user's perspective as they are to decline. When observing these values by category, 58% of the Medical apps improve their rating, while only 34% of the Sports apps improve their mean rating over time.

Our observations suggest that apps that are not performing well have a chance to improve. If we take a more pessimistic view, developers need to proactively work to ensure that their rating levels do not fall. Before we analysed our raw data, we anticipated that the majority of these top apps would have relatively stable ratings and hence did not expect to see high Spearman values (either positive or negative). Our observations of the dynamics of ratings suggest that developers need to be proactive to ensure that the user perceived quality does not decline over time. Although developers may have produced a good quality app, the competitive landscape often pushes users to anchor to higher expectations, making an earlier generation app look dated and hence adding downwards pressure to the mean rating value over time.

4.5. Review Size Evolution

In this section, we present our observations for the final set of research questions that we posed – (i) how do review size distributions change over time and (ii) do apps tend to receive longer or shorter reviews as they age?

In our raw data we had over 17 thousand apps, however once we applied the selection criteria as outlined in Section 3, across all categories, we were able to analyse review size evolution in 7928 apps (both Free and Paid). Of these, we were able to generate valid Spearman correlation coefficients between the Word Count Gini and Age for 64% (5043) of the apps. Slightly fewer Free apps (61%) had valid models compared to Paid apps (66%).

Across all apps with a valid Spearman correlation coefficient, 65% show a positive ρ value. That is, the Word count Gini increases as the app ages. While 35% showed an inverse relationship (see Table 5, and Figure 14 which presents a histogram of the ρ values).

In our analysis, we observed the change in the Gini coefficient of review size over time. A higher Gini coefficient indicates higher inequality, and in this domain this is indicative of a greater proportion of shorter reviews. If apps receive

comparatively longer reviews over time then the wealth (as measured by review size) of the population increases and hence this will be reflected in lower Gini coefficient values. Conversely, when apps continue to receive a proportionally higher number of short reviews, peppered by an occasional long review the inequality within overall population increases which is reflected by a higher Gini coefficient.

Our observations show that in 65% of apps the Gini coefficient is trending upward, as reflected by the positive ρ value. This suggests that as apps age, users tend to leave comparatively shorter reviews. Although this may appear to be undesirable, in our previous work [9] we found that apps tend to get longer reviews when they receive lower ratings and users tend to leave shorter reviews when they rate an app highly. Hence, a good app may improve its mean rating driving an increase in the Gini coefficient (review size) as users leave short reviews – with nothing more than a short word (*e.g.* awesome) indicating their approval. However, developers should perhaps be more concerned if the app receives lower ratings coupled with an increase in Gini coefficient value, since this shows that the users do not like the app and are not providing detailed feedback.

4.6. Threats to Validity

Apple restricts easy access to the App Store reviews data to only on the top 400 ranked apps. Hence, our observations are not contextualized across all rateable App Store apps, nor across all platforms of mobile apps (*e.g.* Android, Blackberry). Additionally, we do not have specific information on how Apple ranks these apps, thus making any direct comparisons with apps from other platforms difficult. Since we ran our last detailed review scrape (20th of July 2012), the App Store has implemented a social voting mechanism for reviews based on criticality, favour-ability, helpfulness and recency, which may have encouraged the growth of reviews post-data collection.

Our study does not account for apps that may actively encourage user feedback via incentive mechanisms or modal dialogs, which may impact review growth. Spam reviews have not been culled from the dataset, as identification of the spam and its impact of growth is not within the scope of this study, although we intend to address this in future work. Anecdotally we manually reviewed a considerable number of reviews and only a very small number appeared to be spam reviews.

We use a very basic metric to analyse review size (character count). The use of different expressive language in reviews; the use of "TXT speak" in reviews and the use of non-English language in a small number of reviews has not been factored into our analysis. Additionally, we have not attempted to factor in different reviewer behaviour. For example, some regularly write short reviews while some regularly write very long ones. We did not attempt to factor in this review size behaviour for the same user reviewing multiple apps.

We rely on commercial statistical software (Stata) for our analysis, allowing us to maintain consistency in computation and presentation for our work. We tested a randomly selected sample of only 500 reviews collected from our scrape script against the data shown by Apple via iTunes in order to verify the reliability of our script and found no defects.

5. Discussion

Our observations pertaining to review growth have several implications for developers. We can now offer empirically informed guidance in the acquisition of app reviews based on our proposed benchmarks.

Regarding review evolution, the significance of the decreasing proportion of large reviews and reduction in user perceived quality may be due to (i) early large reviews being expressive and informative, allowing developers to satisfy user expectations, (ii) general user apathy, possibly due to the large number of alternative apps serving the same or more functionality in a work flow which better aligns with the user's requirements, (iii) an increasingly competitive landscape where competitors are releasing apps for free, at a lower price point, or offering better device support, stability or usability, or (iv) development has focused on resolving issues for an app, but not extending its functionality due to device limitations or the potential higher revenue generated by release of a separate app.

We illustrate the application of our findings with the following scenario: Consider an existing entertainment company that wants to enter the mobile app market to distribute their content and build a new revenue channel. Given the existing brand image and reputation of the company they need a successful flagship app that is well received by the market.

The findings in our work (see Table 3) indicates that a Paid Entertainment app should ideally measure at least 11 reviews after 1 month, 54 after 6, and 110 after 12 months in order to perform above the median for Paid Entertainment apps. This information can be used to set realistic expectations for their development and marketing teams.

Furthermore, as indicated in Table 4, we find that 15.8% of Entertainment apps exhibit super-linear growth which is a target to aim for the app in order to gain attention in the marketplace rapidly. However, the reviews of almost 41% of Entertainment apps do not grow in a generalizable pattern and may be susceptible to spikes in growth driven by new releases, as indicated in Figures 11 and 12. The competitive landscape detailed in Table 5 indicates ρ of 42% positive and 58% negative, suggest that a Paid Entertainment app growth rate is more likely to decline in ratings over time.

Furthermore, Table 5 indicates that 64% of similar apps find a decrease in larger reviews over time. Therefore, as an app ages, the window of opportunity for rich feedback to be obtained from user reviews declines. The combination of this information indicates that the Paid Entertainment app market is volatile with the potential for a high number of app purchases, but with the caveats of a short shelf life if the app is not continuously updated in order to adapt to rapidly evolving user expectations. Our findings show that an app that just stands still in this category will rapidly appear dated. Given the pace of change in the design and feature capabilities of smartphones, apps that do not adjust will receive poor ratings over time since they are compared with newer apps that offer similar features within a more compelling presentation and interaction style.

6. Conclusions and Future Work

Online user reviews in different domains have been investigated for nearly a decade, however there is only a minimal understanding of mobile app user reviews. We have studied approximately 8 million user reviews from the App Store and performed a statistical analysis of these user reviews toward modelling review growth and evolution.

We observed that mobile apps generally receive short reviews, and that both the rating and the category of an application influences the length of a review. Key observations pertaining to growth that we make from this data analysis are that the majority of apps receive well under 50 reviews in their first year, and that the rate of growth for reviews does not present a generalisable pattern with approximately 30% of apps across Free and Paid categories not fitting super-liner, sub-linear, or linear regressions. From our analysis of review evolution over time, we conclude that (i) apps receive a higher number of short reviews in comparison to long reviews as they age, and (ii) around half of the apps decrease in user perceived quality, reflected by the star ratings given to the apps. The relatively short duration within which apps decrease in their perceived quality suggests that the user's expectations are rapidly evolving in this landscape and developers should continuously work towards adapting their apps in order to stay competitive.

Our findings present a benchmark for developers to inform apps for each category and potential growth trends of reviews. Approximately 70% of the review growth fits a trend (linear, sub-linear or super-linear), with remaining 30% exhibiting a non-generalisable growth trend with marked segments of large growth spikes. The review evolution of declining mean ratings and review size proportions also indicate that the early life cycle of apps is where developers gain the most information from reviews.

We currently do not have a strong hypothesis that explains the relationship between an app's category and its influence on the review length beyond postulation. The potential causes for short reviews and why poor ratings tend to elicit longer reviews have yet to be determined in the scope of our current experiment. Furthermore, our visual analysis of the non-fitting 30% of app growth plots allude to apparent growth patterns that invite closer temporal-based inspection. We are interested in exploring how certain positive or negative sentiments used in reviews for an app early on may impact future ratings, and if they correlate with ratings and review size over time.

These gaps grant us opportunities for future work where we analyse the actual content of the review using opinion mining and text analysis approaches. Given our finding that most reviews tend to be short and about the size of a tweet, we are hopeful that the text analysis techniques developed for analysing Twitter streams can be applied and hence we intend to explore that arc in the near future.

Acknowledgments

We would like to thank Rosmarie Schneider, Milica Stojmenovic, Scott Barnett and Maheswaree Kissoon Curumsing for their time and comments on earlier drafts of this work.

References

[1] I. E. Vermeulen and D. Seegers, "Tried and Tested: The Impact of Online Hotel Reviews on Consumer Consideration," *Tourism Management*, vol. 30, pp. 123–127, Feb. 2009.

[2] X. Yu, Y. Liu, J. X. Huang, and A. An, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, pp. 720–734, Apr. 2012.

[3] J. A. Chevalier and D. Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, vol. 43, no. 3, pp. 345–354, 2006.

[4] J. Leino and K.-J. Räihä, "Case Amazon: Ratings and Reviews as Part of Recommendations," in *Proceedings of the 2007 ACM Conference on Recommender Systems*, pp. 137–140, ACM, 2007.

[5] W. Duan, B. Gu, and A. B. Whinston, "The Dynamics of Online Word-of-Mouth and Product Sales – An Empirical Investigation of the Movie Industry," *Journal of Retailing*, vol. 84, pp. 233–242, June 2008.

[6] 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.

[7] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Micro-blogging as Online Word of Mouth Branding," in *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 3859–3864, ACM, 2009.

[8] J. Gebauer, Y. Tang, and C. Baimai, "User Requirements of Mobile Technology: Results from a Content Analysis of User Reviews," *Information Systems and E–Business Management*, vol. 6, pp. 361–384, 2008.

[9] R. Vasa, L. Hoon, K. Mouzakis, and A. Noguchi, "A Preliminary Analysis of Mobile App User Reviews," in *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pp. 241–244, Nov. 2012.

[10] M. Harman, Y. Jia, and Y. Zhang, "App Store Mining and Analysis: MSR for App Stores," in *Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on*, pp. 108–111, June 2012.

[11] H.-N. Chen and C.-Y. Huang, "An Investigation into Online Reviewers Behavior," *European Journal of Marketing*, vol. 47, no. 10, pp. 10–10, 2013.

[12] E. Platzer, "Opportunities of Automated Motive-Based User Review Analysis in the Context of Mobile App Acceptance," in *Proceedings of the 22nd Central European Conference on Information and Intelligent Systems*, pp. 309–316, Sept. 2011.

[13] D. Godes and D. Mayzlin, "Using online conversations to study word-of-mouth communication," *Marketing Science*, vol. 23, no. 4, pp. 545–560, 2004.

[14] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[15] S. J. Koyani, R. W. Bailey, J. R. Nall, S. Allison, C. Mulligan, K. Bailey, M. Tolson, and N. C. Institute, *Research-Based Web Design & Usability Guidelines*. National Cancer Institute, 2004.

[16] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, pp. 319–340, September 1989.

[17] F. T. C. Tan and R. Vasa, "Toward a social media usage policy," in *22nd Australasian Conference on Information Systems Proceedings*, pp. 84–89, TeX Users Group, November-December 2011.

[18] V. Kostakos, "Is the crowd's wisdom biased? a quantitative analysis of three online communities," in *International Conference on Computational Science and Engineering Proceedings*, vol. 4, pp. 251–255, August 2009.

[19] 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. Information Technology and Systems in the Internet–Era.

[20] R. Chandy and H. Gu, "Identifying Spam in the iOS App Store," in *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, WebQuality '12, pp. 56–59, ACM, Apr. 2012.

[21] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Summarizing Microblogs Automatically," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, (Stroudsburg, PA, USA), pp. 685–688, Association for Computational Linguistics, 2010.

[22] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Experiments in Microblog Summarization," in *Proceedings of IEEE Second International Conference on Social Computing*, 2010.

[23] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-Tracking and the Dynamics of the News Cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 497–506, ACM, 2009.

[24] Apple Inc., "App Store Tops 40 Billion Downloads with Almost Half in 2012." http://bit.ly/VNWHHs, 2013.

[25] R. Vasa, *Growth and Change Dynamics in Open Source Software Systems*. PhD thesis, Swinburne University of Technology, 2010.

[26] W. Zucchini, "An Introduction to Model Selection," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 41–61, 2000.

[27] N. Fenton and S. L. Pfleeger, *Software Metrics: A Rigorous and Practical Approach*. London, UK: International Thomson Computer Press, second ed., 1996.

[28] C. Gini, "Measurement of Inequality of Incomes," *The Economic Journal*, vol. 31, pp. 124–126, Mar. 1921.

[29] R. Vasa, M. Lumpe, P. Branch, and O. Nierstrasz, "Comparative Analysis of Evolving Software Systems Using the Gini Coefficient," in *Proceedings of the 25th IEEE International Conference on Software Maintenance (ICSM'09)*, IEEE Computer Society, 2009.

[30] L. Hoon, R. Vasa, J.-G. Schneider, and K. Mouzakis, "A Preliminary Analysis of Vocabulary in Mobile App User Reviews," in *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pp. 245–248, Nov. 2012.

[31] M. M. Lehman, D. E. Perry, J. C. F. Ramil, W. M. Turski, and P. Wernik, "Metrics and Laws of Software Evolution – The Nineties View," in *Proceedings of the Fourth International Symposium on Software Metrics (Metrics '97)*, pp. 20–32, Nov. 1997.