DOI:10.1145/3141771

The varying review dynamics seen in different app stores can help guide future app development strategies.

BY STUART MCILROY, WEIYI SHANG, NASIR ALI, AND AHMED E. HASSAN

User Reviews of Top Mobile Apps in Apple and Google App Stores

ONE OF THE unique aspects of app stores is the convenience of providing user feedback.[13] Users can effortlessly leave a review and a rating for an app, providing quick feedback for developers. Developers are then better able to update their apps. This feedback mechanism contrasts with traditional feedback mechanisms like bug-reporting systems (such as Bugzilla), which are negative in nature, as they include only bugs, unlike reviews, which can be positive. Moreover, reviews can even serve as a means for deriving additional app requirements.[7]

Developers of top apps might be overwhelmed by the large number of received reviews. Several papers (such as by Fu et al.,[5] Galvis Carreño and Winbladh,[6] and Google Analytics[7]) and commercial efforts (such as Applause Analytics[3]) have proposed solutions to help developers cope with large numbers of reviews.

A 2013 study of reviews of iOS apps by Pagano and Maalej[20] found that on average a free app receives 37 reviews per day, while paid apps receive approximately seven reviews per day,[20] and another study of iOS apps found that 50% of studied free apps receive only 50 reviews in their first year.[11] Yet no prior research examined the reviews in the Google Play store, considering, say, "Is the data normally distributed or highly skewed, with only a small number of apps receiving a substantial number of reviews on a daily basis?"

Here, we explore the question of how pervasive are the frequently reviewed apps in the Google Play store. In particular, we empirically cover app reviews from the perspective of the developers of the top apps there. Through an analysis of reviews for the top 10,713 apps in the Google Play store over a period of two months—January 1 to March 2, 2014—we found:

More than 500 reviews daily. Only 0.19% of the studied apps received more than 500 reviews per day;

Majority of studied apps. Almost 88% of the studied apps received only a small number (20 or fewer) reviews per day; and

Correlates with reviews. The number of downloads and releases correlated with the number of received reviews, while the app category did not play a major role.

Some of our observations differ from other studies of user reviews of iOS

» key insights

- **The characteristics of user reviews differ depending on app store.**
- **Few mobile apps in the Google Play store attract large numbers of user reviews.**
- **More app downloads and releases correlate with more reviews in the Google Play store, whereas app category plays only a minor role.**

IMAGE BY ANDRIJ BORYS ASSOCIATES/SHUTTERSTOCK

apps,[11] highlighting the need for additional in-depth investigation of the reviewing dynamics in both stores.

Mobile App Analytics

A Vision Mobile survey of 7,000 developers, also in 2014, found 40% of them made use of user-analytics tools and 18% used crash-reporting and bug-tracking tools. Other studies also found that developers need tools for app analytics. For example, a 2013 study by Pagano and Bruegge[19] of how feedback occurs following initial release of a software product identified the need to structure and analyze feedback, particularly when it involves a large amount of feedback.

A number of app-analytics companies, including App Annie,[1] specialize in tools designed to help developers understand how users interact with their apps, how developers can help generate revenue (such as through in-app purchases, e-commerce, and direct buy), and how to leverage user demographics of the apps. These companies also provide developers overviews of user feedback and crash reports. Google promotes its own extensive analytics tools for Android developers as a key competitive differentiator relative to other mobile app stores. The tools measure how users use an app (such as by identifying user locations and how users reached the app). They also track sales data (such as how developers generate revenue through in-app purchases and the effect of promotions on app sales[2]). However, other than crash-reporting tools, many analytics tools today are mostly sales-oriented rather than software-quality-oriented involving bugs, performance, and reliability.

Other studies have highlighted the effect of reviews of mobile apps on an app's success.[9,15,19] Harman et al.[9] found a strong correlation between app ratings and an app's total download numbers. User reviews include information that could help developers improve the quality of their apps and increase their revenue. Kim et al.[15] interviewed app buyers, finding reviews are a key determinant in their decisions to purchase an app. A survey by Lim et al.[16] found reviews are one of the top reasons for users to choose an app. Likewise, Mudambi et al.[18] showed that user reviews have a significant effect on sales of online products.

The importance of user reviews has motivated many studies, as well as our own work analyzing and summarizing user reviews for mobile apps (see Table 1). Iacob and Harrison[12] built a rule-based automated tool to extract feature requests from user reviews of mobile apps, an approach that identifies whether or not a user review contains a feature request. Chandy and Gu[3] identified spam reviews in the Apple (iOS) App Store, using a technique that achieved high accuracy with both labeled and unlabeled datasets. Carreño and Winbladh[6] used opinion-mining techniques and topic modeling to successfully extract requirements changes from user reviews. Fu et al.[5] introduced an approach for discovering inconsistencies in apps, analyzing the negative reviews of apps through topic analysis to identify reasons for users liking or disliking a given app. Khalid et al.[14] manually analyzed and categorized one- and two-star reviews, identifying the issues (such as the hidden cost of using an app) about which users complained. Chen et al.[4] proposed the most extensive summarization approach to date, removing uninformative reviews and prioritizing the most informative reviews before presenting a visualization of the content of reviews. Guzman and Maalej[8] performed natural language processing techniques to identify app features in the reviews and leveraged sentiment analysis to identify whether users like such features. Our own work differs from these studies, as it aims to provide context about when the other techniques would be needed.

Pagano and Maalej[20] and Hoon et al.[11] analyzed the content of reviews of both free and paid apps in the Apple App Store, answering a similar research question as ours about the number of received reviews, but there are major differences between them and us in terms of findings, methodologies, and context, or Android vs. iOS (see Table 2).

Studied Apps

Martin et al.[17] noted that not all stores provide access to all their reviews, leading to biased findings when studying reviews. To avoid such bias, we collected all reviews on a daily basis, ensuring we would include all available reviewers. However, the Google Play Store provides access to only the 500 latest reviews for an app. If more than 500 reviews are received in the 24-hour period between daily runs of our crawler, then the crawler does not collect those reviews. This limitation means we thus offer a conservative estimate of the number of reviews for apps that receive more than 500 reviews per 24-hour time period. We based our Google Play store crawler on an open source crawler called the Akdeniz Google Play crawler (https://github.com/Akdeniz/google-playcrawler) to extract app information (such as app name, user ratings, and reviews). Running it meant we were simulating a mobile device over approximately two months—January 1 to March 2, 2014.

We collected review information from 12,000 free-to-download apps from the Google Play store. From among 30 different categories, including photography, sports, and education, we selected the top apps in each category in the U.S. based on app-analytic company Distimo's (acquired by App Annie) ranking of apps for a total of 12,000; Distimo ranked the top 400 apps for each of the 30 categories. We used Distimo's Spring 2013 list of top apps. Of the 12,000 top apps, 1,287 were not accessible during our two-month crawl because some of them might have been removed from the store. We thus collected data from 10,713 top apps, with a total of 11,047 different releases during the studied time period.

Our own selection of top apps might have biased our results, possibly generalizing to only the top, stable, free apps in the Google Play store. Nevertheless, we studied successful apps we felt were more likely to have a large user base and receive a large number of reviews, rather than blindly study all apps. We chose apps that had been popular one year before we began our study because we were

Table 1. Our observations on Google Play apps compared to the Pagano and Maalej[20] and Hoon et al.[11] observations on the Apple (iOS) App Store.

Item	In the Apple App Store, from Pagano and Maalej[21]	In the Apple App Store, from Hoon et al.[11]	In the Google Play Store, from us	Notes
Reviews received	Average of 22 reviews per day, with 36.87 for free apps and 7.18 for paid apps	Median of 50 reviews in first year for free apps and 30 reviews in first year for paid apps	Average of seven reviews per day, with median of no reviews per day for free apps	We found fewer average and median user reviews compared to Pagano and Maalej[21] and more user reviews than Hoon et al.[11] Reviews were skewed, with median number of received reviews at 0 and 88% of the studied apps receiving 20 reviews or fewer per day.
Number of reviews received	Facebook received 4,275 reviews in one day	(not studied)	Only 0.19% of apps received more than 500 reviews, and the top 100 most-reviewed apps had 6,000 to 43,000 reviews in the two-month study period.	Pagano and Maalej[21] were the first to observe that some apps (for them, the Facebook app) might receive a large number of reviews per day. We were first to explore this observation—apps receiving a large number of reviews per day—in depth, finding that while some apps might receive a large number of reviews, only 0.19% of all studied apps received more than 500 reviews per day. Most top apps might not benefit much from automated approaches that leverage sophisticated techniques (such as topic modeling) given the small number of reviews they received and their limited length.
Effect of app category	Number of daily reviews differs by category	Certain categories receive greater numbers of reviewers than others	No relation	Compared to both iOS studies, we found no relation between an app's category and number of received reviews, once we controlled for number of downloads and number of releases.
Spike in number of reviews decreases following release	Number of reviews decreases over time following a release	(not studied)	The standard deviation of received reviews deviates from the median directly following release and returns back to normal afterward.	Both stores showed evidence of spikes in number of reviews immediately following a new release.
Average length of a review	Average of 106 characters and median of 61 characters	Average of 117 characters and median of 69 characters	Average of 64 characters and median of 36 characters	Reviews in the Google Play Store were shorter than in the Apple App Store. Median length of reviews demonstrated that the distribution of review length is highly skewed, with long reviews as outliers.

interested in stable, mature apps that had not been released within the past few months to avoid the expected burst of reviews following an app's initial release.[20] We focused on free-to-download apps, since recent work showed that free apps receive five times as many reviews as paid apps.[20] Moreover, over 90% of downloaded apps were, at the time, of the free-to-download variety, according to Gartner. Such apps use other revenue models (such as freemium, in-app purchases, and ads). The developers of such apps are thus concerned about the effect of reviews on their revenue.[9]

Findings

Here, we present our findings, as in Table 2, concerning the reviews from the Google Play store while comparing our results with prior studies.

Number of received reviews. On the number of received reviews in the Google Play Store

Finding 1. Most apps (88% of those of the 10,713 we studied) received few reviews during our studied time period. The average and median number of reviews were fewer than Pagano's and Maalej[20] and greater than Hoon et al.;[11]

Finding 2. The number of user reviews were skewed; similar findings were reported by Pagano and Maalej;[20] and

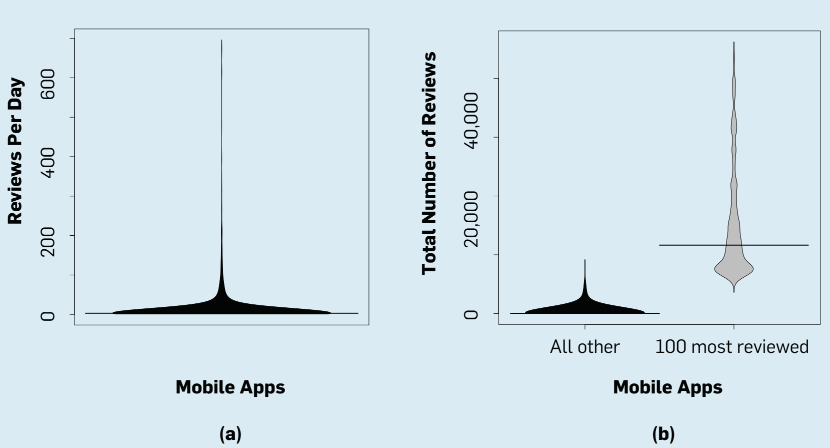
Implication. Most top apps might not benefit much from automated approaches to analyzing reviews that leverage sophisticated techniques (such as topic modeling) given the small number of received user reviews and their limited length.

We plotted the number of reviews per day, as well as total number of received reviews, using a beanplot combining a boxplot with a kernel-density-estimation function. Figure 1a reports the median number of reviews per day was 0. We found 20, or 0.19%, of the 10,713 studied apps received 500 or more reviews; as mentioned earlier, 500 would be a conservative estimate, whereas 88% of the apps in our 10,713-app dataset received fewer than 20 reviews per day. Additionally, the median total number of reviews was 0 during the study period. We also calculated the number of words in each of the received reviews, with median number of words per review at 46.

Table 2. Datasets of prior work mining reviews of mobile apps.

Paper	App Store	Apps	Reviews
Iacob and Harrison[12]	Google Play Store	161	3,279
Galvis and Carreno[7]	Google Play Store	2	710
Fu et al.[6]	Google Play Store	171,493	13,286,706
Chen et al.[5]	Google Play Store	4	169,097
Pagano and Maalej[21]	Apple App Store	1,100	1,126,453
Hoon et al.[11]	Apple App Store	17,000	8,700,000

Figure 1. Beanplots showing number of reviews per day and in total.



We found fewer average reviews per day than Pagano and Maalej[20] possibly due to any of several factors. The first is we collected reviews from stable top apps that had been released for at least one year, whereas Pagano and Maalej[20] may have collected new apps and not focused on top apps. The second was that our estimates for the frequently reviewed apps were conservative; we did not count more than 500 reviews in a day. For instance, Pagano and Maalej reported that Facebook received 4,275 reviews in a day, with such large numbers increasing the overall reported average number of received reviews on a daily basis. We separated the apps into two groups: 100 most-reviewed apps and all other apps. Figure 1b reports there was a large gap in the total number of reviews among the 100 most-reviewed apps. The total number of reviews of the 100 most-reviewed apps ranged from 43,000 to 6,000 in the two-month study period. The reviews themselves were short, much shorter (approximately 40%) than the reviews in the Apple App Store. We also observed a notable skew in the length of reviews in both stores.

Influence of app category and downloads on number of reviews. In the Google Play Store

Finding 3. The number of downloads and releases correlated with the number of received reviews, whereas an app's category did not play a major role during the study period. On the other hand, Pagano and Maalej[20] and Hoon et al.[11] both reported a relation between an app's category and the number of received reviews; and

Implication. The relationship between number of received reviews and an app's category should be explored further, especially in light of the discrepancy between the two app stores.

Here, we investigate the effect of an app's number of downloads, number of releases, and app category on the number of received reviews. We built a regression model with an app's number of received reviews as the dependent variable. Due to the notable skew in the number of reviews, we log-transformed the number of reviews before building the linear-regression model.

Figure 2 plots the total number of reviews using the built-regression model. We included three plots, each keeping the median values of the other factors

Figure 2. Plots of the total number of reviews (logged) on the *y*-axis and three separate graphs of app categories, number of downloads, and number of releases on the *x*-axis; the graphs reflect the relation between the three factors and the total number of reviews.

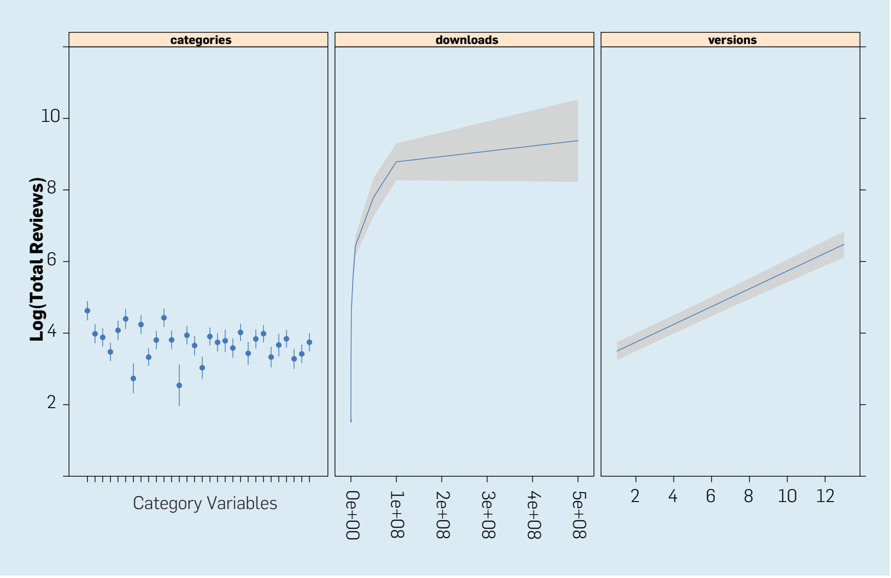


Figure 3. A nomograph of the effect of new releases, app category, and number of downloads on total number of reviews received.

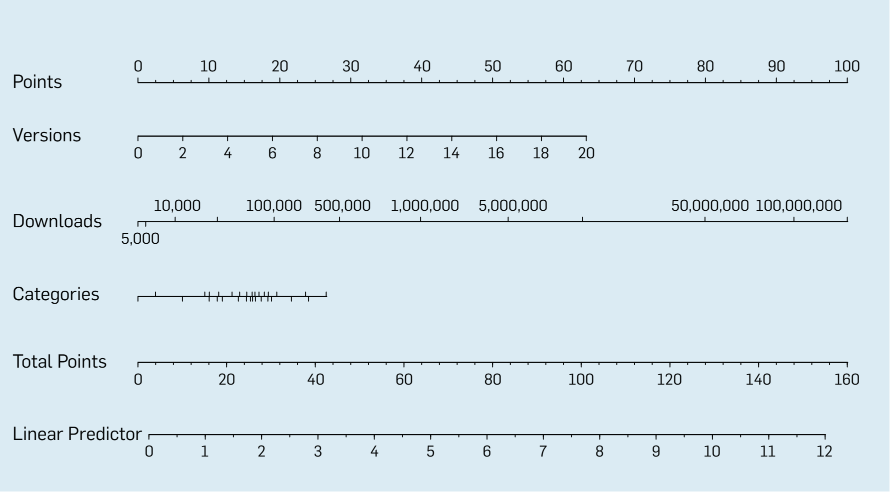
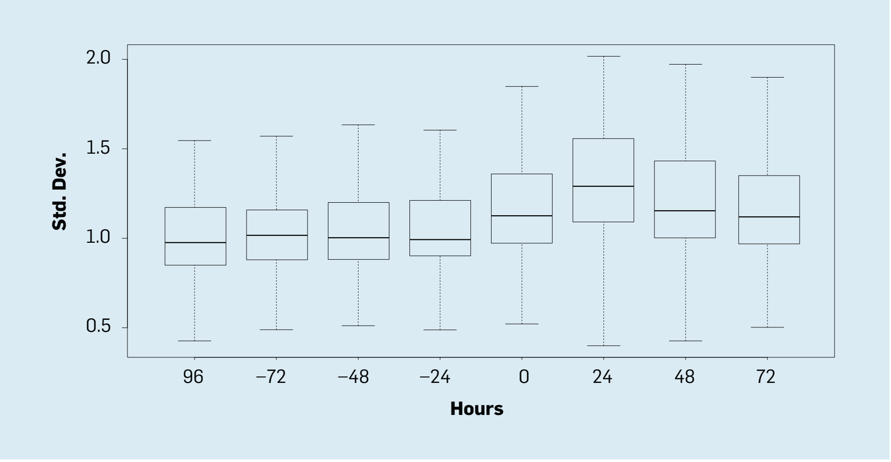


Figure 4. Standard deviation of new reviews every 24 hours before and after the first collected release for each studied app; each boxplot represents the standard deviation from the median number of reviews for each app at that time.



the same so we could see how each factor affects the total number of reviews.[10] The gray bands around the plotted lines are bootstrap confidence intervals for our estimates.

We generated a nomogram (see Figure 3) to visualize the results of our regression model,[10] helping us examine the effect of each factor while controlling for other factors. The nomogram consists of a series of scales. The Linear Predictor scale is the total number of reviews in log scale. To calculate the total number of reviews, we can draw a straight line from the value of the "total points" scale to the linear predictor scale. The total points are calculated by summing the points of each of the scales of the three factors: releases, downloads, and categories. To calculate the points value of each factor, we can draw a line from the value in the factor scale to the points scale. The value in the points scale becomes the points for that factor. For example, releases = 2, downloads = 100,000, and categories = tools. We found that 2-releases corresponded to approximately seven points, 100,000-downloads corresponded to approximately 20 points, and the tools category corresponded to approximately five points. The sum was 32 total points, which corresponded to approximately 2.5 log scale, or 316 total user reviews.

We found that as the number of downloads and releases increased, the total number of reviews also increased. We found no relation between individual categories (such as communications, social, tools, and review count) when we controlled for number of downloads and releases. In contrast, Pagano and Maalej[20] and Hoon et al.[11] observed a relation between categories and number of received reviews in the Apple App Store; however, neither study controlled for the other metrics in its analysis. Those studies observed a relation between categories and number of reviews that may be due to the interaction between categories and number of downloads or between categories and number of releases.

Spike in reviews following a release. Finally, concerning the spike in reviews following a release of an app in Google Play Store

Finding 4. Both the Google Play store and the Apple App Store showed evidence of a spike in reviews following a release; and

Implication. Greater effort examining user reviews should follow a release in order to improve app quality.

Pagano and Maalej[20] reported that the number of received reviews decreased over time after a release, suggesting releases contribute to new reviews. We observed the same kind of correlation for the Google Play store. Figure 4 outlines a boxplot of the median number of reviews for each studied app across each of its releases, showing a spike in reviews directly on and after an app's release day.

However, still not clear is if these spikes were due to an app attracting new users following its release or to current users becoming more inclined to review the app. Looking closer at our nomogram, we note that many releases (more than 20) for an app has as much of an effect as an app with 10 million downloads. Frequent releases thus ensure an app's user base is more engaged as it begins providing feedback.

Frequent releases ensure an app's user base is more engaged as it begins providing feedback.

Conclusion

A very small percentage of the top apps we studied (0.19% of 10,713) have ever received more than 500 reviews per day, yet most studied apps received only a few reviews per day. The number of received reviews for the studied apps did not vary due to the category to which the app belonged, varying instead based on number of downloads and releases. Some of our results highlight differences between the Google Play store and the Apple App Store.

Additional studies are needed to better understand the review dynamics across both stores. Researchers should thus examine whether other empirical findings hold across them. In particular, techniques designed to assist mobile-app developers should be optimized for each store.

References

1. App Annie Analytics; http://www.appannie.com/app-store-analytics/
2. Applause; https://www.applause.com/testing/
3. Chandy, R. and Gu, H. Identifying spam in the iOS app store. In *Proceedings of the Second Joint WICOW/AIRWeb Workshop on Web Quality* (Lyon, France, Apr. 16). ACM Press, New York, 2012, 56–59.
4. Chen, N., Lin, J., Hoi, S,C.H., Xiao, X., and Zhang, B. AR-Miner: Mining informative reviews for developers from the mobile app marketplace. In *Proceedings of the 36th International Conference on Software Engineering* (Hyderabad, India, May 31–June 7). ACM Press, New York, 2014, 767–778.
5. Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., and Sadeh, N. Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, IL, Aug. 11–14). ACM Press, New York, 2013, 1276–1284.
6. Galvis Carreño, L.V. and Winbladh, K. Analysis of user comments: An approach for software requirements evolution. In *Proceedings of the 2013 International Conference on Software Engineering* (San Francisco, CA, May 18–26). IEEE Press, Piscataway, NJ, 2013, 582–591.
7. Google Analytics; http://www.google.ca/analytics/mobile/
8. Guzman, E. and Maalej, W. How do users like this feature? A fine-grained sentiment analysis of app reviews. In *Proceedings of the 22nd International Requirements Engineering Conference* (Karlskrona, Sweden, Aug. 25–29). IEEE Press, Piscataway, NJ, 2014, 153–162.
9. Harman, M., Jia, Y., and Zhang, Y. App store mining and analysis: MSR for app stores. In *Proceedings of the Ninth Working Conference on Mining Software Repositories* (Zurich, Switzerland, June 2–3). Piscataway, NJ, 2012.
10. Harrell, F.E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer, New York, 2001.
11. Hoon, L., Vasa, R., Schneider, J.-G., Grundy, J. et al. *An Analysis of the Mobile App Review Landscape: Trends and Implications.* Technical Report. Swinburne University of Technology, Faculty of Information and Communication Technologies, Melbourne, Australia, 2013.
12. Iacob, C. and Harrison, R. Retrieving and analyzing mobile apps feature requests from online reviews. In *Proceedings of the 10th International Workshop on Mining Software Repositories* (San Francisco, CA, May 18–19). IEEE Press, Piscataway, NJ, 2013, 41–44.
13. Johns, T. *Replying to User Reviews on Google Play.* Android Developers Blog, June 21, 2012; http://android-developers.blogspot.ca/2012/06/replying-to-user-reviews-on-google-play.html
14. Khalid, H., Shihab, E., Nagappan, M., and Hassan, A. What do mobile app users complain about? *IEEE Software 32*, 3 (May-June 2015), 70–77.
15. Kim, H.-W., Lee, H.L., and Son, J.E. An exploratory study on the determinants of smartphone app purchase. In *Proceedings of the 11th International DSI Decision Sciences Institute and 16th APDSI Asia Pacific Region of Decision Sciences Institute Joint Meeting* (Taipei, Taiwan, July 12–16, 2011).
16. Lim, S.L., Bentley, P.J., Kanakam, N., Ishikawa, F., and Honiden, S. Investigating country differences in mobile app user behavior and challenges for software engineering. *IEEE Transactions on Software Engineering 41*, 1 (Jan. 2015), 40–64.
17. Martin, W., Harman, M., Jia, Y., Sarro, F., and Zhang, Y. The app-sampling problem for app store mining. In *Proceedings of the 12th Working Conference on Mining Software Repositories* (Florence, Italy, May 16–17). IEEE Press, Piscataway, NJ, 2015.
18. Mudambi, S.M. and Schu, D. What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly 34*, 1 (2010), 185–200.
19. Pagano, D. and Bruegge, B. User involvement in software evolution practice: A case study. In *Proceedings of the 2013 International Conference on Software Engineering* (San Francisco, May 18–26). IEEE Press, Piscataway, NJ, 2013, 953–962.
20. Pagano, D. and Maalej, W. User feedback in the App Store: An empirical study. In *Proceedings of the 21st IEEE International Requirements Engineering Conference* (Rio de Janeiro, Brazil, July 15–19). IEEE, Piscataway, NJ, 2013.

Stuart Mcilroy (mcilroy@cs.queensu.ca) is a Ph.D. student at Dalhousie University, Halifax, Canada.

Weiyi Shang (shang@encs.concordia.ca) is an assistant professor and Concordia University Research Chair in Ultra-Large-Scale Systems in the Department of Computer Science and Software Engineering at Concordia University, Montreal, Canada.

Nasir Ali (cnali@memphis.edu) is an assistant research professor at the University of Memphis, Memphis, TN.

Ahmed E. Hassan (ahmed@cs.queensu.ca) is Canada Research Chair in Software Analytics and NSERC/BlackBerry Software Engineering Chair in the School of Computing at Queen's University, Kingston, Canada.