Consolidated Review of

Rise of the Planet of the Apps: A Systematic Study of the Mobile App Ecosystem

1. Strengths:

New perspective analyzing detailed download statistics from mobile app markets. Interesting analysis of app developer income & competing revenue strategies. Novel findings on app clustering effect.

Generally well-done and in-depth analysis - - Well written. Nice analysis, nice insights and nice model.

2. Weaknesses

The data on which the paper is based is not great. The data studied is limited to 4 Android appstores that do not include the most popular Google Play, so it's difficult to infer how generalizable the conclusions are. The paid app dataset is fairly small (~5k apps). Unsure the results are generalizable.

Unclear what the implications are for networks, content distribution, etc. Tenuous fit for IMC. Some metrics could be chosen better (see below)

3. Comments

The paper is very well presented and structured. It consists of an analysis of four third-party app markets for Android. I like the fact that the paper does not only describe the analytical results and the graphs but goes in depth in trying to find reasons for the exhibited behavior, it creates metrics and models to explain and frame these characteristics and has nice discussion sections on how the results can be useful for (although I have reservations about this section).

There are some assumptions they made that I feel the authors could have justified better. (Also a few pages didn't render properly so maybe I missed something) the lack of the play store seems like a pretty big omission. However Google has made it harder lately to crawl the market so maybe they had no choice. In particular installing third party apps requires more technical knowledge than some users might have and means users know how to install pirated apps which probably skews the benefits of paid apps a bit. Also using comments as a substitute for app downloads also seems like a big assumption. I think some validation of those assumptions, even a smaller scale study of the play store, would make the work a lot stronger. On the other hand the findings presented in the paper are very consistent with the "common wisdom" discussed in the developer community - the findings aren't implausible, and I think this work would be useful for a lot of people.

I believe that this is a solid and well-done measurement paper on an important topic for mobile app developers. I found the finding on the clustering effect especially interesting and found the analysis to be convincing (if not entirely bullet proof). My main concern with this paper is that it does not layout what the implications are for networks, content distribution, etc., the general audience for IMC. The results on app pricing and income are interesting in their own right, but I'm not sure IMC is the appropriate audience for them. I would suggest adding some implications of the clustering effect on content placement and caching. A minor nitpick: In Sec 4.1, your metric for determining how many users focus on a few categories doesn't seem to take into account that some users will post much more than others. So, a user that posts on 2 apps in the same category will be counted the same as a user that posts on 30 apps in the same category. Since most users seem to post on a very small number of apps (paper cites 99% < 30), perhaps you should weight the metric based on the number of apps commented on (e.g., categories/comments) and see how confident you are of the mean of this value across users as comments increases. I suspect there is a lot of noise in the value for users that only post a few times

I have a number of reservations over the claims of this paper:

- ❖ I am not sure that the CDF (Figure 5a) is a good way to identify the clustering effect because it also incorporates the distribution of the number of comments per user. For example, if most users have less than two comments they can only fall in the first two bins of the CDF. Another example: In the caption of figure 5a, you say that 94% of the users only commend on up to 5 categories (and therefore only 6% voted in more than 5 categories)… However, what percentage of the users has more than 5 comments in general? Is it significantly higher than 6% of the users? How close is this line to the distribution of number of comments per user? I think this graph only makes sense if it also includes also the CDF on the number of comments per user (which should hopefully fall significantly below the distribution of comments in unique categories). Additionally, another parameter that might affect these results is the distribution of downloads per category. Some categories are more likely to have a large number of downloads: in particular, games and entertainment apps are much more likely to be downloaded rather than tools and photography and education apps. And therefore, some categories are much more likely to attract comments (as they have more downloads). Does the distribution of number of downloads per category affect the CDF shown in figure 5? In other words, does the fact that categories have disproportionate number of apps/downloads affect the distribution of number of categories that users commented on?
- ❖ Similarly, for the Affinity metric you define, does not take account the probability of an app in a category being downloaded (it only takes into account the number of apps in the category).
- ❖ Another reservation about this work is that although it is considering four different markets it is not considering the Google Play market, which is the dominant one. The authors refer to ref [33] to justify their choice indicating that [33] shows that third-party markets are a "significant" proportion of the of the app ecosystem. I looked at [33] and figure 8 does not appear to be that strongly supporting this claim...so it would be nice if some more clear indication of whether the third party markets considered include an indeed significant portion. The about current number of apps on the Google

Play is about 700K so your numbers from table 1 are definitely smaller.

- ❖ The justification of the flattening of the distribution for small x in figure 3 is that it approximates the number of users: could you strengthen your argument by comparing with the number of users in your markets? I do not think you do this.
- ❖ The other aspect that worries me about this paper is analysis of the app pricing and income, which is done only on one market, as this was the only one with paid apps. I think some of the conclusions on revenues are not strongly supported: why would developers put apps on a smaller market rather than on Google Play which would give it much more visibility? I have of course no way of proving this as there is no data but I would encourage you perhaps to weaken your claims in this section considering this bias...I would argue that most paid apps appear on Google Play and not in third-party markets. I may be wrong...
- ❖ Does the selected market place affects your result? For example, do people trust SlideMe for credit card information etc.? Or do they prefer to buy apps from the original Google Play store or Amazon? Are there any indications that the distribution in number of downloads between free vs. Paid apps is similar to other markets?
- ❖ This leads to the question about the model proposed in Section 5.1.
  - o How do you tune the download prob p for the apps in the same category?
  - o For users who download app from several categories, how do you choose which of these categories this app is drawn from with prob. p? Uniformly from existing categories, or always download from the same category as previous step?
  - o Can the distribution of downloads per user, d, be somehow estimated based on the public user comments from Anzhi market, instead of using just the mean number of downloads per user?
  - o If the users' interests can be captured (based on which categories most frequently visited), then the first step of the model may be enhanced (instead of choosing an app randomly based on ZG).
- ❖ Section 6.1 Figure 12, top figure: it is not very clear that the number of downloads is negatively correlated with the price.
- ❖ Section 6.3, in comparing the two possible revenue strategies: paid apps vs. free app with advertisement, the paper tries to estimate the necessary ad income to achieve the same revenue of paid app. While this is reasonable, lumping the downloading and pricing statistics across all genres of applications doesn't seem to be reasonable. For instance, gaming apps tend to be more complex and cost more to develop. Pricing for gaming apps will be very different from other categories, and the download count will also differ across genre. Why not perform the same analysis for different categories of apps?

Regarding the practical implications:

- ❖ Although there is a list at the conclusion section I am still not sure what are the important practical implications of this work: More specifically: "The existence of locality in user Downloads… can help appstores design efficient caching mechanisms that will improve the speed of delivering apps to end users"
- ❖ Usually caching mechanisms in the back end take into account multiple users. How does knowing that a single user is slightly more likely to download an app from the same category can help? Even if we assume that we can pre-cache on his phone other apps, we don't know which app from the same category…
- ❖ "The understanding of download patterns, like clustering effect, can help appstores to design better recommendation systems" Why is that more helpful than the current state of the art that offers recommendations based on: "users like you that have downloaded X also downloaded Y". Does the knowledge that users download apps from the same category offer any significant benefit?
- ❖ "Our model of app downloads can be used by appstores to estimate future app popularity and downloads. This will enable appstores to pinpoint problematic apps and either favor them through better recommendations or remove them from the market. " Can they use this model to make any prediction about a specific app? How? Why should they delete apps if they are not popular?
- ❖ "Understanding which pricing models result in larger revenue can help developers to choose an appropriate pricing policy for their apps to increase app popularity and their income" In practice advertisements use a very complicated auction system and it is hard to predict revenue. This might also depend on the popularity of the app and the number of impressions that are given to specific apps.
- ❖ The paper seems to choose to completely ignore the fact that users' interest drive the app. downloads, and will inherently visit certain app. categories (more app. downloads from that genre) more frequently than others. The way the hypothesis (B) is phrased in Section 4 makes it sound like if a user, by chance, downloads an app from a clusters, he/she is more likely to download another app. from the same cluster, and completely ignores the underlying cause, which is users' interests.

In essence I like the style of this paper, its aims, models, and most of its findings, I am however worried about the fact that the type of data on which the analysis was done weakens the contributions and therefore the final claims.

4. Summary from PC Discussion

The PC liked this paper's analysis and findings. However, we had concerns about how representativeness of the three markets compared with the Play store and the App store (e.g., due to a biased user and app population). As much as possible, we would like the authors to validate and discuss how that their claims would apply to the larger app markets.

5. Authors' Response

We thank the IMC reviewers for their insightful reviews and their valuable feedback for improving our paper. We identified three areas that the majority of the comments focus on: (i) the clustering effect property, (ii) the pricing study, and (iii) the possible practical implications of this work. We have tried to address these concerns in the final version of our paper:

1. Regarding clustering effect, there were concerns about the study proving the existence of the proposed effect on user downloads based on user comments. To this end, we extended our analysis and we found that although 20% of the users ade more than five comments, the percentage of users that commented on more than five different categories is just 6%, which is significantly lower. Moreover, we showed that there is

no dominant category in terms of downloads, so the popularity distribution of app categories does not seem to have a considerable influence on user downloads. We have addressed these comments by providing Figures 5(a) and 5(d) and discussing the results in Section 4.1. Moreover, there were two questions regarding our proposed model: (i) how we tuned the model parameters, and (ii) how we selected an app category for clustering-based downloads. We explained the methodology we followed to tune the model parameters in a way that they result in the best data fit, i.e., minimum distance from actual data, and we described how we randomly choose an app category from the previous downloads for each simulated clustering-based download. We have added a more detailed description of our model in Section 5.1 and we explain how we tuned the model parameters in Section 5.2.1.

2. Concerning the comments on app pricing analysis, some of them focus on the small sample of paid apps hosted in SlideMe and on the trustworthiness of this marketplace, e.g., for payments. We agree that these are valid concerns, and we have added a discussion in the first paragraph of Section 6 to explain why developers actually choose third-party app stores like SlideMe to host their paid apps. One reason is that they place their apps in many different app stores at the same time in order to gain more popularity or attract users with devices that have pre-installed third-party app stores like SlideMe. In Section 2 we list several vendors that indeed provide devices with the SlideMe app store pre-installed. Other reasons are Google's regional restrictions and financial benefits offered by these app stores. Another concern was the representativeness and significance of the app stores we analyzed. To this end, we have enriched Section 2.3 to quantify the significance of our dataset: we show that at the last day of our data collection we had 52.7% of the total apps and 35% of the total downloads, comparing with Google Play app population and downloads the same time period. Thus, we believe that we analyzed a significant portion of the market. To address the question on how the necessary ad income for free apps with advertisements changes for different app categories, we have extended our analysis to measure the necessary ad income per each different app category. The results are presented in Figure 18 and discussed at the end of Section 6.3.

3. Regarding the possible implications of our findings, we have added a new separate section (Section 7). In this section we explain how a typical LRU cache can improve the app delivery performance, and we show the negative effect of clustering-based downloads on the LRU cache hit ratio, comparing with ZIPF-based downloads through simulations. These results are shown in Figure 19. Thus, we conclude that new replacement policies are needed to deal with clustering effect. In the same section, we discuss a few ideas on how our study can help app stores to design better recommendation systems. Moreover we briefly discuss a few other possible implications, such as effective prefetching, improving app popularity, and maximizing developers' income.

We have also incorporated other fixes and suggestions from the reviews.