EMSE manuscript No.
(will be inserted by the editor)

Studying the Consistency of Star Ratings and Reviews of Popular Free Hybrid Android and iOS Apps

Hanyang Hu · Shaowei Wang · Cor-Paul Bezemer · Ahmed E. Hassan

Received: date / Accepted: date

Abstract Nowadays, many developers make their mobile apps available on multiple platforms (e.g., Android and iOS). However, maintaining several versions of a cross-platform app that is built natively (i.e., using platform-specific tools) is a complicated task. Instead, developers can choose to use hybrid development tools, such as PhoneGap, to build hybrid apps. Hybrid apps are based on a single codebase across platforms. There exist two ways to use a hybrid development tool to build a hybrid app that runs on multiple platforms: (1) using web technologies (i.e., HTML, Javascript and CSS) and (2) in a common language, which is then converted to native code.

We study whether these hybrid development tools achieve their main purpose: delivering an app that is perceived similarly by users across platforms. Prior studies show that users refer to star ratings and user reviews, when deciding to download an app. Given the importance of star ratings and user reviews, we study whether the usage of a hybrid development tool assists app developers in achieving consistency in the star ratings and user reviews across multiple platforms.

We study 68 hybrid app-pairs, i.e., apps that exist both in the Google Play store and Apple App store. We find that 33 out of 68 hybrid apps do not receive consistent star ratings across platforms. We run Twitter-LDA on user reviews and find that the star ratings of the reviews that discuss the same topic could be up to three times as high across platforms. Our findings suggest that while hybrid apps are better at providing consistent star ratings and user reviews when compared to cross-platform apps that are built natively, hybrid apps do not guarantee such consistency. Hence, developers should not solely rely on hybrid development tools to achieve consistency in the star ratings and reviews that are given by users of their apps. In particular, developers should track closely the ratings and reviews of their apps across platforms, so that they can act accordingly when platform-specific issues arise.

Keywords Mobile apps; star rating; user reviews; Twitter-LDA

Hanyang Hu, Shaowei Wang (corresponding author), Cor-Paul Bezemer, Ahmed E. Hassan
Queen's University, Kingston, ON, Canada
E-mail: {hyhu, shaowei, bezemer, ahmed}@cs.queensu.ca

1 Introduction

To offer their mobile apps across platforms (e.g., Android and iOS), developers build their apps using native (i.e., using platform-specific tools) or hybrid development tools. Apps that are created using native development tools are referred to as native apps. Since creating cross-platform apps using native development tools requires a deep understanding of different programming languages and frameworks, hybrid development tools are becoming popular among developers. Apps that are created using hybrid development tools are a subset of cross-platform apps and are commonly referred to as hybrid apps. In 2014, approximately 5% of the apps in the Google Play Store were hybrid apps [47].

Developers use hybrid development tools to save on development costs. A small to medium-sized cross-platform development project costs $35,000 to $70,000 using native development kits, but only $20,000 to $40,000 using hybrid development tools [46]. Also, in contrast to developing apps using native development tools, hybrid development tools allow developers to work on a single codebase, from which corresponding apps on each supported platform are generated. There are two types of hybrid apps: (1) hybrid apps that are built using HTML, CSS and JavaScript (e.g., using the PhoneGap framework [1]) and (2) hybrid apps that are built in a common language such as C#, and then converted to native code for each supported platform by a hybrid development framework (e.g., using the Xamarin framework [28]). In both cases, the hybrid app has the same code base across platforms: in the first case, the HTML content is displayed in an internal browser (a webview) on each platform. In the second case, the apps share the code base from which the native code is generated.

Maintaining a consistent user experience is one of the most valued tasks for cross-platform developers, as observed by Joorabchi et al [19]. Developers try to make their apps look and behave similarly across the platforms on which the apps are available. Whether developers succeed in delivering a consistent look and behavior, is partially reflected from the star ratings and user reviews of the apps. In the case of hybrid apps, part of the effort of ensuring the consistency of the apps is done by the development tools. Thus, the selection of development tools, either native or hybrid, could potentially influence the consistency of star ratings and user reviews.

In our previous work [17], we showed that cross-platform apps that are developed using native tools achieve a low consistency of star ratings and user reviews. Little is known about the influence of hybrid development tools on the consistency of star ratings and user reviews. In this paper, we explore in depth the consistency of the star ratings and user reviews of 68 hybrid apps across the Android and iOS platforms. In particular, we address the following research questions:

RQ1 How consistently do users rate hybrid apps across platforms?
32 out of 68 hybrid apps receive a significantly different distribution of star ratings across both studied platforms. For 13 of these apps, the effect size of the difference between the studied platforms ranges from small to large. Compared to our previous study on native cross-platform apps from the top 50 charts, the studied hybrid apps have more consistent overall star ratings across platforms.

RQ2 How consistent are the raised topics in reviews rated across platforms? We use Twitter-LDA to extract the most important issues (topics) that are raised in user reviews. For the same review topic, star ratings can be three times as high on one of the platforms. 144 out of 424 of the identified topics are not equally rated across platforms. 93 of those 144 topics receive higher average star ratings in Android. Similar to the findings in our previous work [17], we find that users appear to show a preference for hybrid apps on certain platforms. Developers can analyze the differences in user ratings for reviews in which the same issues are raised, to better understand the priorities and encountered issues of users across platforms.

This paper is further organized as follows. Section 2 discusses related work. Section 3 presents the design of our empirical study. The findings of our study are presented in Section 4. In Section 5, we discuss the implications of our findings. Section 6 addresses the threats to the validity of our study. Finally, Section 7 presents our conclusion.

2 Related Work

Cross-platform apps and hybrid apps have recently been popular subjects among software engineering researchers. First, we discuss related work on hybrid apps and second, we discuss related work on native cross-platform apps. Finally, we give a brief overview of related work on app review analysis.

2.1 Hybrid Apps

Smutny [42] discusses the advantages of apps built using HTML5. He shows 4 characteristics of hybrid development tools, including PhoneGap, and outlines 9 of such tools as examples. Ohrt and Turau [31] compare 9 hybrid development tools and outline their advantages and drawbacks. The majority of the related work on hybrid apps focuses on comparing the software quality (e.g., memory usage) of hybrid apps with respect to different hybrid development tools.

Dalmasso et al. [8] refer to hybrid development tools as "write once run anywhere tools". They compare several popular hybrid development tools, such as PhoneGap and Titanium. They find that apps developed by PhoneGap consume the least memory, CPU and power. A similar study by Heikotter et al. [16] also shows that PhoneGap is the most viable hybrid development tool in terms of the look and feel of PhoneGap apps. Palmieri et al. [33] compare application programming interfaces, programming languages, and supported platforms of four hybrid development tools. They conclude that Rhodes stands out from other hybrid development tools because of its support for web-based services and the MVC framework.

Viennot et al. [47] created the PlayDrone dataset that has over 1.1 million Android apps. Viennot et al. identify 59,354 hybrid apps that are built using PhoneGap, Adobe Air, or Titanium. Ali and Mesbah [3] use a tool named ClassyShark to inspect class names of 1.1 million Android apps in the PlayDrone dataset. Their study focuses on identifying hybrid apps. Ali and Mesbah identify 15k hybrid apps from 1.1 million Android apps.

Our study is different from most of the related work because we focus on analyzing star ratings and user reviews whereas previous studies focus on the hybrid development tools or the characteristics of the software artifacts that are produced by such hybrid development tools.

2.2 Cross-Platform Apps

Our previous study [17] examines the 19 cross-platform apps that are in the top 50 apps charts of the Google Play Store and the App Store. Users from 14 out of 19 cross-platform apps give inconsistent star ratings and user reviews. However, all 19 apps in our prior study are developed natively. In this paper, we focus on hybrid apps. Man et al. [23] developed CrossMiner which captures 7 app issues from user reviews of cross-platform apps on Google Play, App Store and Windows Store. They discover that cross-platform apps for different platforms generate different issue distributions. They also find that issues related to "crash" and "network" raise more user concerns compared to other types of issues. Joorabchi et al. [20] implement a tool named CHECKCAMP to detect inconsistencies in cross-platform apps. Joorabchi et al. find that functionality, data, layout and style are the four most pervasive types of inconsistency in cross-platform apps. In another study, Joorabchi el al. [19] survey and interview mobile app developers. They find that maintaining the behavioural consistency is an important task for developers of cross-platform apps. The behavioural consistency stated in Joorabchi et al.'s study includes offering the same functionality and a similar look-and-feel across different platforms.

2.3 Analysis of Mobile App Ratings and Reviews

The analysis of mobile app ratings and reviews has received a great deal of attention over the past years. In this section, we give a brief overview of the most important related work. For a more thorough overview of studies on analyzing star ratings and user reviews, we refer to Martin et al.'s survey [27].

2.3.1 Studies Using Manual Analysis

The first direction within work on app review analysis focuses on manual analysis of app reviews. Manual analysis of app reviews can yield in-depth insights on the behaviour and preferences of the users of an app. For example, Pagano and Maalej [32] discovered that most user feedback is given shortly after a new app release. Hassan et al. [15] studied emergency updates of top Android apps, and found that such updates are often due to simple mistakes. Khalid et al. [21] studied complaints in app reviews and found that users complain the most about functional errors, and that they are most critical of privacy and ethics-related issues.

2.3.2 Studies Using Automated Analysis

The second direction within work on app review analysis focuses on automated analysis of app reviews and ratings. Automated analysis of app reviews allows a

much larger number of reviews to be analyzed at once, and can yield insights into store-wide behaviour.

Several studies were done on the relation between app rating and other app-related factors. For example, Harman et al. [14] mined 32,108 BlackBerry apps and found a strong correlation between the average star rating of an app and the number of downloads. Tian et al. [45] identified that install size, the number of promotional images and the target SDK version are the most related factors to app rating. Martin et al. [26] found that higher priced releases are more likely to have a positive impact on the success of an app. Noei et al. [30] found that some device attributes (such as CPU) have a stronger relation with the user-perceived quality than other app attributes (such as the number of UI inputs).

Several studies were done on automatically analyzing the contents of app reviews, often using Latent Dirichlet Allocation (LDA). For example, Chen et al. [7] developed "AR-Miner", a tool which collects app reviews, analyzes them using LDA and searches for reviews that provide valuable information. WisCom [10] uses LDA to detect why people like or dislike an app. Guzman and Maalej [13] combined LDA with sentiment analysis and proposed an automated approach for extracting reviews that contain feature-related information, which can be used for requirements evolution. Gu and Kim [12] automatically summarize reviews using natural language processing. Vu et al. [49] allow developers to search for relevant reviews using keywords.

Automated analysis of app reviews is also used to guide the evolution of apps. For example, Palomba et al. [34] showed how to link user reviews to source code changes, and found that in most cases, developers follow-up on requests or complaints in user reviews when updating their apps. Villarroel et al. [48] proposed an automated approach for release planning based on the information that is available in app reviews, such as bug reports and feature requests. Panichella et al. [36] used a combination of natural language processing, sentiment analysis and text analysis to automatically classify reviews into categories, with the goal of improving the app maintenance and evolution process. Di Sorbo et al. [9] automatically generate user summaries of app reviews. Palomba et al. [35] proposed an approach that recommends source code changes based on user reviews.

In our work, we conduct automated analysis of app reviews. However, our work is different from prior work as we focus on hybrid apps, in particular, inconsistencies of such apps across platforms.

3 Empirical Study Design

In this section, we describe the design of our empirical study of star ratings and user reviews of hybrid apps. Figure 1 gives an overview of the steps of our study. In the remainder of this section, we describe our data collection process in more detail.

3.1 Identifying Hybrid Apps

We focus on apps from the Google Play Store and App Store due to their dominant market share [40]. In particular, we focus on popular apps in the U.S. version of

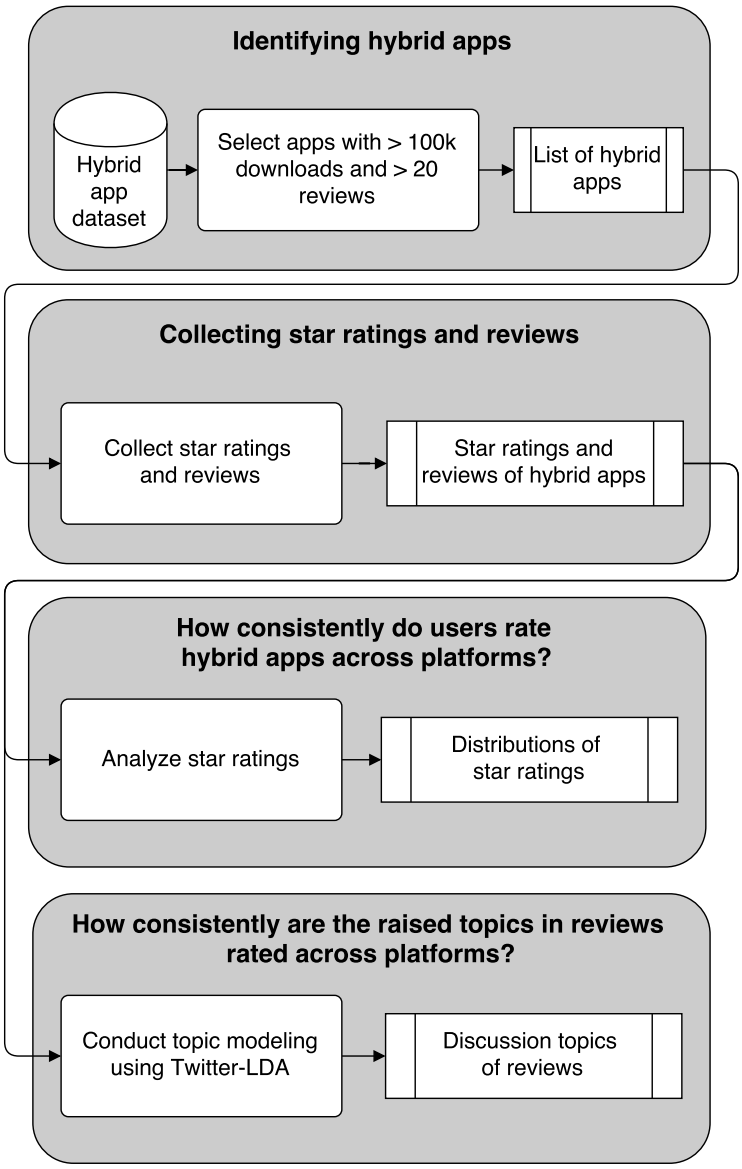

Fig. 1: Overview of the empirical study design.

the stores, as our servers are located in North America and they have access to the North American version of the stores only. We study free apps instead of paid apps due to the fact that free apps account for more than 90% of the total app downloads [37]. In addition, when studying free apps, it is not necessary to account for differences in app prices across platforms, which may influence expectations and opinions of users.

We use Ali and Mesbah's public dataset of hybrid apps [3] to get an initial selection of hybrid app-pairs. However, the majority of the hybrid apps in this dataset are non-popular "zombie apps". For example, the median number of downloads is 100 for these apps and the median number of ratings is 4. Therefore, we limit our study to hybrid apps that have at least 100,000 downloads in the Google Play store (the App store does not provide the exact number of downloads of each app) and at least 20 reviews on each platform. As a result, we end up with 251 hybrid app-pairs from Ali and Mesbah's dataset. The names and means of identification across app stores of the studied apps can be found online[1].

3.2 Collecting Star Ratings and Reviews

After we identify 251 hybrid apps, we collect their star ratings and user reviews. Neither the App Store nor the Google Play Store provide public APIs to collect the entire set of reviews for apps. However, given an app ID, such as 284882215 for the *Facebook* app, Apple offers a public RSS (Rich Site Summary) feed that allows us to collect the 500 most recent star ratings and reviews for the app [4]. We used a crawler for the Google Play Store [2] to collect the star ratings and reviews of Android apps. Martin et al. [25] warned about using reviews that are collected during a short time frame for review analysis when doing long-term research studies. Therefore, we crawl the user reviews of the studied hybrid app pairs in both stores for a 3-month period (May 5 2017 – August 4 2017). Hence, for apps with less than 500 reviews in total, we crawl all reviews and for apps with more than 500 reviews in total, we crawl the 500 latest reviews plus all reviews that were posted during the 3-month period. We remove all duplicate reviews.

A small number of star ratings and reviews may introduce bias in the analysis of star ratings and reviews. Hence, we exclude 183 hybrid apps from the 251 identified hybrid apps because the number of collected reviews is smaller than 20 in either the Google Play store or Apple's App store. Finally, 68 hybrid apps remain as the main subjects of our study. Table 1 shows the number of collected reviews for the studied hybrid apps. The median number of collected reviews for a hybrid app is 519 for the Android version of the app and 267 for the iOS version of the app.

4 Empirical Study Results

In this section, we present the results of our empirical study. We discuss the motivation, approach and results for each of our research questions.

4.1 How Consistently Do Users Rate Hybrid Apps Across Platforms?

Motivation: A 2015 survey shows that 69% of the app users consider app rating an important or very important factor when deciding whether to download an app or not [39]. In addition, 77% of the app users will not download an app that is rated lower than 3-stars. Since a larger number of app downloads usually generates more

[1] `https://zenodo.org/record/1181881`

App Name	# of Collected Reviews		
	Android	iOS	Total
2048 by Gabriele Cirulli	515	301	816
3D CAD Models Engineering	262	20	282
ACTPhoto	500	255	755
AlAhliMobile	639	362	1,001
AT&T Mark the Spot	512	505	1,017
Between - Private Couples App	1,468	1,321	2,789
BMO Harris Mobile Banking	500	502	1,002
Bridge Base Online	337	105	442
Bubble Cat Adventures	155	28	183
Companion for Dota 2 Lite	512	55	567
Cookie Jam	2,356	596	2,952
D3LTA	316	32	348
Das Erste	115	29	144
DC Metro and Bus	514	503	1,017
Discover Mobile	3,198	1,292	4,490
Don't Look Back	523	500	1,023
Doomsday Preppers	248	525	773
Eldhelm - Free Online CCG/RPG	502	56	558
ElfYourself by OfficeMax	390	510	900
Fat Fingers: for eBay Bargains	176	20	196
FXCM Trading Station Mobile	510	73	583
GameStop	764	506	1,270
Gangster Paradise	560	112	672
Gartic	625	25	650
GoGoVan	529	32	561
Hard Time (Prison Sim)	1,946	503	2,449
HealthTap for U.S. Doctors	320	158	478
Horse Academy	500	92	592
Hungry Howies Pizza	1,101	166	1,267
Ikariam Mobile	529	176	705
IKEA Catalog	637	564	1,201
Inflation RPG	614	507	1,121
Learn German - 3400 words	240	75	315
Let's Fish: Sport Fishing	1,297	607	1,904
Linksys Smart Wi-Fi	1,305	653	1,958
McPixel Lite	501	134	635
Meine Bank	336	62	398
Michaels Stores	752	536	1,288
Michigan Lottery Mobile	520	121	641
Minnesota State Fair	499	169	668
Music Inc	503	49	552
MVG FahrInfo Mnchen	159	33	192
MyHeritage	664	551	1,215
NetBenefits	464	294	758
Odd Socks	558	508	1,066
Ohio Lottery	507	279	786
Panda Jam	523	503	1,026
Paychex Mobile	519	310	829
Pepi Bath Lite	521	103	624
Perkd	241	20	261
PhotoMania - Photo Effects	504	93	597
Popscene (Music Industry Sim)	835	186	1,021
PULL&BEAR	366	57	423
Rollercoaster Mania	559	513	1,072
Safe Pregnancy and Birth	42	25	67
Schwan's Food Delivery	519	185	704
ShopYourWay	551	501	1,052
SkipTouch	463	116	579
Stardoll Access	562	484	1,046
Stick Tennis	861	576	1,437
Take Control of the Tower	520	35	555
The Tower	529	548	1,077
Urban Rivals	132	520	652
Walgreens	4,806	1,983	6,789
Warframe Nexus	867	484	1,351
Weight Watchers Mobile	4,305	4,684	8,989
Wrestling Revolution 3D	9,254	49	9,303
ZIP Code Tools	215	437	652
Median	519	267	765.5

Table 1: The number of collected reviews for the 68 studied hybrid apps.

revenue for free apps (through showing advertisements within the app), developers of hybrid apps can benefit from high star ratings of their apps on all platforms. In this study, we are interested in whether hybrid apps on different platforms receive consistent star ratings. For each studied hybrid app, we analyze the consistency of two distributions of star ratings across the two studied platforms. Inconsistent distributions of star ratings indicate a difference in software quality. As Joorabchi et al. point out in their survey [19], achieving consistency in software quality is one important task for app developers. Additionally, having a bad app on one platform can discourage users of other platforms from downloading the app, for example, through negative word of mouth advertising. Therefore, developers of hybrid apps that have inconsistent star ratings across platforms need to consult user reviews to better understand the differences in the perceived software quality of their app across platforms.

Approach: We analyze the star ratings at both the platform level and the app level. In particular, we first analyze the distribution of all collected star ratings at the platform level. We then analyze the distribution of all star ratings at the app level. We are interested in the consistency of the distribution of the star rating at both levels.

We compare the distribution of the star ratings for both versions of a hybrid app by using the average star rating, the skewness and the kurtosis [18] of both distributions. A small number of reviews of an app could bias the results of the study.

We first calculate the average star rating for each studied hybrid app across both studied platforms. We use the Mann-Whitney U test [24] to test the significance level of the difference of means between the two distributions since that test does not require a normal distribution. We use the Bonferroni correction to adjust the default significance level of the Mann-Whitney U test, which is $\alpha = 0.05$. Specifically, to decide whether the star ratings for the iOS and Android-version of the same app differ significantly, we perform a Mann-Whitney U test with a significance level $\alpha = 0.0007$ (0.05/68).

In order to quantify the difference in star ratings, we calculate Cliff's delta (δ) [22] effect size on the two distributions of star ratings in Android and iOS. The range of Cliff's delta δ is [-1, 1]. The absolute value of the Cliff's delta δ is used to quantify the differences in two distributions. In particular, we use the following thresholds for interpreting δ [41]:

$$\text{Effect size} = \begin{cases} negligible(N), & \text{if } |\delta| \leq 0.147. \\ small(S), & \text{if } 0.147 < |\delta| \leq 0.33. \\ medium(M), & \text{if } 0.33 < |\delta| \leq 0.474. \\ large(L), & \text{if } 0.474 < |\delta| \leq 1. \end{cases}$$

Consistent distributions of star ratings across platforms will have a p-value that is larger than 0.0007 in the Mann-Whitney U test, which suggests that both Android users and iOS users rate the app similarly. On the other hand, inconsistent distributions of star ratings across platforms will have a p-value that is smaller than 0.0007 in the Mann-Whitney U test. The effect size (negligible, small, medium or large) quantifies the difference between the distributions.

We calculate the skewness and kurtosis for both studied versions of each studied hybrid app. The skewness of a distribution captures the level of symmetry in terms

App Name	App Type	Average Rating			M.W.[2]	Effect[3] Size
		Android	iOS	Ratio[1]		
2048 by Gabriele Cirulli	Games	3.8	4.6	0.8	Y	S
3D CAD Models Engineering	Catalogs	4.4	4.5	1.0	N	-
ACTPhoto	Education	1.2	2.0	0.6	Y	N
AlAhliMobile	Finance	3.3	2.3	1.5	Y	S
AT&T Mark the Spot	Utilities	2.0	1.6	1.2	Y	N
Between - Private Couples App	Social Networking	4.4	4.8	0.9	Y	N
BMO Harris Mobile Banking	Finance	1.9	1.7	1.2	Y	N
Bridge Base Online	Games	3.5	2.2	1.6	Y	S
Bubble Cat Adventures	Games	3.6	2.6	1.4	N	-
Companion for Dota 2 Lite	Entertainment	4.0	4.1	1.0	N	-
Cookie Jam	Games	4.4	4.1	1.1	Y	S
D3LTA	Photo & Video	3.5	4.1	0.8	N	-
Das Erste	Entertainment	2.5	2.2	1.2	N	-
DC Metro and Bus	Navigation	3.6	3.5	1.0	N	-
Discover Mobile	Finance	4.7	4.8	1.0	Y	N
Don't Look Back	Games	3.2	4.2	0.8	Y	N
Doomsday Preppers	Games	4.2	4.3	1.0	N	-
Eldhelm - Free Online CCG/RPG	Games	3.9	4.0	1.0	N	-
ElfYourself by OfficeMax	Entertainment	4.6	3.0	1.5	Y	S
Fat Fingers: for eBay Bargains	Lifestyle	2.6	2.8	1.0	N	-
FXCM Trading Station Mobile	Finance	3.6	3.0	1.2	N	-
GameStop	Entertainment	3.8	2.2	1.8	Y	S
Gangster Paradise	Games	4.5	4.5	1.0	N	-
Gartic	Games	4.6	2.5	1.8	Y	M
GoGoVan	Travel	4.5	4.3	1.0	N	-
Hard Time (Prison Sim)	Games	4.5	3.8	1.2	Y	N
HealthTap for U.S. Doctors	Medical	4.4	4.4	1.0	N	-
Horse Academy	Games	2.4	2.3	1.0	N	-
Hungry Howies Pizza	Lifestyle	4.3	4.4	1.0	N	-
Ikariam Mobile	Games	2.6	2.6	1.0	N	-
IKEA Catalog	Lifestyle	2.2	1.7	1.3	Y	N
Inflation RPG	Games	4.5	4.7	1.0	N	-
Learn German - 3400 words	Education	4.5	4.5	1.0	N	-
Let's Fish: Sport Fishing	Games	4.4	4.4	1.0	N	-
Linksys Smart Wi-Fi	Utilities	4.3	4.4	1.0	Y	N
McPixel Lite	Games	3.8	4.2	0.9	N	-
Meine Bank	Finance	2.7	2.4	1.1	N	-
Michaels Stores	Lifestyle	4.3	4.1	1.1	N	-
Michigan Lottery Mobile	Entertainment	3.3	2.3	1.4	Y	N
Minnesota State Fair	Entertainment	2.4	3.3	0.7	Y	S
Music Inc	Games	3.5	3.5	1.0	N	-
MVG FahrInfo Mnchen	Travel	2.1	2.5	0.9	N	-
MyHeritage	Reference	4.0	3.6	1.1	Y	N
NetBenefits	Finance	2.0	2.2	0.9	N	-
Odd Socks	Games	4.1	4.2	1.0	N	-
Ohio Lottery	Reference	2.0	1.9	1.1	N	-
Panda Jam	Games	4.1	3.2	1.3	Y	N
Paychex Mobile	Business	1.8	1.9	0.9	N	-
Pepi Bath Lite	Education	3.7	3.0	1.2	Y	N
Perkd	Lifestyle	3.3	2.4	1.4	N	-
PhotoMania - Photo Effects	Photo & Video	4.6	4.7	1.0	N	-
Popscene (Music Industry Sim)	Games	4.2	3.7	1.1	Y	N
PULL&BEAR	Lifestyle	3.0	2.1	1.4	Y	S
Rollercoaster Mania	Games	4.2	3.4	1.2	Y	N
Safe Pregnancy and Birth	Medical	4.2	4.6	0.9	N	-
Schwan's Food Delivery	Food & Drink	3.9	2.8	1.4	Y	S
ShopYourWay	Lifestyle	3.6	2.7	1.3	Y	S
SkipTouch	Games	2.8	2.2	1.3	Y	N
Stardoll Access	Social Networking	2.9	2.7	1.1	N	-
Stick Tennis	Games	4.1	3.9	1.1	Y	N
Take Control of the Tower	Games	3.8	2.9	1.3	Y	N
The Tower	Games	4.6	4.5	1.0	N	-
Urban Rivals	Games	4.1	3.9	1.1	N	-
Walgreens	Lifestyle	4.4	4.0	1.1	Y	S
Warframe Nexus	Reference	3.1	3.0	1.1	N	-
Weight Watchers Mobile	Health & Fitness	4.5	4.5	1.0	N	-
Wrestling Revolution 3D	Games	4.5	3.7	1.2	Y	S
ZIP Code Tools	Reference	3.8	4.7	0.8	Y	N
Values across all apps		3.0	3.1	1.0	Y	N

[1]Ratios in this and following tables are calculated by Android / iOS.
[2]Mann-Whitney's U test: Y: p-value smaller than 0.0007. N: otherwise
[3]Effect size: N: negligible, S: small, M: medium, L: large

Table 2: Statistics of the star ratings of hybrid apps

App Name	Skewness			Kurtosis		
	Android	iOS	Ratio[1]	Android	iOS	Ratio
2048 by Gabriele Cirulli	-0.8	-2.7	0.3	-0.7	7.3	-0.1
3D CAD Models Engineering	-2.0	-1.6	1.2	3.3	1.3	2.5
ACTPhoto	3.8	1.1	3.4	13.6	-0.7	-20.5
AlAhliMobile	-0.3	0.7	-0.5	-1.7	-1.2	1.3
AT&T Mark the Spot	1.2	1.8	0.7	0.0	2.1	0.0
Between - Private Couples App	-2.0	-3.6	0.6	3.4	14.1	0.2
BMO Harris Mobile Banking	1.3	1.8	0.7	0.1	1.8	0.1
Bridge Base Online	-0.5	0.9	-0.6	-1.3	-0.8	1.7
Bubble Cat Adventures	-0.6	0.4	-1.4	-1.4	-1.5	0.9
Companion for Dota 2 Lite	-1.2	-1.3	0.9	0.4	0.6	0.6
Cookie Jam	-1.9	-1.3	1.5	2.2	0.1	33.1
D3LTA	-0.4	-1.4	0.3	-1.4	0.9	-1.6
Das Erste	0.5	0.9	0.5	-1.4	-1.0	1.5
DC Metro and Bus	-0.7	-0.5	1.2	-1.1	-1.3	0.9
Discover Mobile	-3.1	-4.1	0.8	9.4	16.5	0.6
Don't Look Back	-0.1	-1.5	0.1	-1.4	0.7	-2.1
Doomsday Preppers	-1.5	-1.4	1.1	1.6	1.8	0.9
Eldhelm - Free Online CCG/RPG	-1.0	-1.2	0.8	-0.6	0.0	-16.1
ElfYourself by OfficeMax	-2.9	0.0	73.0	7.5	-1.7	-4.3
Fat Fingers: for eBay Bargains	0.4	0.3	1.2	-1.7	-1.7	1.0
FXCM Trading Station Mobile	-0.6	0.0	56.2	-1.2	-1.8	0.6
GameStop	-0.9	0.9	-0.9	-0.9	-0.6	1.5
Gangster Paradise	-2.4	-2.5	0.9	4.3	5.1	0.9
Gartic	-2.6	0.5	-5.1	5.4	-1.6	-3.3
GoGoVan	-2.4	-1.7	1.4	4.1	1.4	2.9
Hard Time (Prison Sim)	-2.2	-1.0	2.3	3.7	-0.5	-7.7
HealthTap for U.S. Doctors	-2.1	-1.9	1.1	3.3	2.4	1.3
Horse Academy	0.6	0.7	0.8	-1.4	-1.0	1.4
Hungry Howies Pizza	-1.6	-2.1	0.8	1.6	2.9	0.5
Ikariam Mobile	0.4	0.4	1.0	-1.3	-1.4	1.0
IKEA Catalog	0.9	1.8	0.5	-0.9	1.7	-0.5
Inflation RPG	-2.3	-3	0.8	4.5	9.5	0.5
Learn German - 3400 words	-1.9	-2.0	1.0	4.6	3.7	1.2
Let's Fish: Sport Fishing	-2.0	-1.8	1.1	3.5	3.1	1.1
Linksys Smart Wi-Fi	-1.8	-2.2	0.8	1.9	3.9	0.5
McPixel Lite	-0.9	-1.4	0.7	-0.9	0.2	-3.9
Meine Bank	0.3	0.7	0.5	-1.6	-1.0	1.5
Michaels Stores	-1.7	-1.3	1.3	2.0	0.3	6.5
Michigan Lottery Mobile	-0.3	0.7	-0.4	-1.7	-1.1	1.6
Minnesota State Fair	0.6	-0.3	-1.9	-1.0	-1.5	0.7
Music Inc	-0.5	-0.5	1.0	-0.9	-1.1	0.8
MVG FahrInfo Mnchen	1.0	0.6	1.7	-0.5	-1.4	0.4
MyHeritage	-1.1	-0.7	1.7	-0.3	-1.2	0.2
NetBenefits	1.1	0.8	1.3	-0.1	-1.0	0.1
Odd Socks	-1.4	-1.4	1.0	0.4	1.0	0.4
Ohio Lottery	1.1	1.3	0.9	-0.3	0.3	-0.8
Panda Jam	-1.3	-0.2	6.6	0.2	-1.4	-0.1
Paychex Mobile	1.5	1.3	1.2	0.7	-0.1	-6.2
Pepi Bath Lite	-0.8	0.0	-961.2	-1.2	-1.5	0.8
Perkd	-0.4	0.2	-1.7	-1.4	-1.2	1.2
PhotoMania - Photo Effects	-2.8	-3.1	0.9	8.7	10.2	0.9
Popscene (Music Industry Sim)	-1.5	-0.7	2.1	1.0	-1.0	-1.1
PULL&BEAR	0.0	1.0	0.0	-1.7	-0.7	2.3
Rollercoaster Mania	-1.5	-0.5	3.1	1.0	-1.1	-0.9
Safe Pregnancy and Birth	-1.5	-2.5	0.6	0.5	4.9	0.1
Schwan's Food Delivery	-1.1	0.3	-4.1	-0.6	-1.7	0.3
ShopYourWay	-0.6	0.3	-2.2	-1.4	-1.8	0.8
SkipTouch	0.2	0.9	0.2	-1.5	-0.4	4.2
Stardoll Access	0.1	0.2	0.6	-1.8	-1.3	1.4
Stick Tennis	-1.4	-1.0	1.4	0.6	-0.4	-1.5
Take Control of the Tower	-0.9	0.0	81.4	-0.9	-1.4	0.6
The Tower	-2.6	-2.3	1.1	6.0	4.2	1.4
Urban Rivals	-1.3	-1.0	1.3	0.3	-0.6	-0.5
Walgreens	-1.9	-1.2	1.6	2.6	-0.3	-9.8
Warframe Nexus	-0.1	0.0	4.9	-1.3	-1.3	1.0
Weight Watchers Mobile	-2.3	-2.4	1.0	5.3	5.8	0.9
Wrestling Revolution 3D	-2.4	-0.7	3.3	5.4	-1.0	-5.5
ZIP Code Tools	-0.9	-3.4	0.3	-0.8	10.1	-0.1

[1]Ratios in this and following tables are calculated by Android / iOS.

Table 3: Skewness and kurtosis ratio for the distribution of star ratings of the studied hybrid apps

of mean and median, i.e., the skewness of the distribution of star ratings represents how positive or negative users feel about that version of the app. A negative skew means that users feel negative (i.e., more lower star ratings) about the app, while a positive skew means that users feel positive (i.e., more higher ratings). While there is no official threshold for skewness, a skew smaller than -1 or larger than 1 means that the skew is substantial [11].

Kurtosis explains the peakedness of a distribution. The Gaussian distribution has a kurtosis of 0. Note that we use the excess kurtosis throughout this paper [29]. A positive kurtosis means that the distribution has a higher peak than the Gaussian distribution, while a negative kurtosis means that the distribution is flatter. A positive kurtosis means that users have a relatively strong consensus on the average star rating of the app, while a negative kurtosis means that there is no clear consensus (i.e, agreement) between the users. Table 3 shows the skewness and kurtosis for ratings of the studied hybrid apps. In addition, Table 3 shows the ratio of the skewness and kurtosis across platforms to easily identify the apps that have the largest differences across platforms. If the ratio is 1, the skewness/kurtosis is the same across platforms. If the ratio is < 1, the iOS ratings have a larger skew/kurtosis. If the ratio is > 1, the Android ratings have a larger skew/kurtosis.

Ideally, the ratings of an app have a high skew (i.e., users are positive) and high kurtosis (i.e., users give similar ratings to the app). A low kurtosis indicates that users feel very differently about the app, leaving room for possible improvements to make users across platforms perceive the app equally. Hence, developers can study the kurtosis of the ratings of their apps to identify possible avenues for improvements.

Results: **32 out of 68 (47%) hybrid apps have a significantly different distribution of star ratings. 24 of these 32 apps have a higher average star rating in Android.** For 11 out of these 32 hybrid apps, the effect size of the difference ranges from small to large, as shown in Table 2, indicating that for these 11 apps the difference is noticeable in practice. In total, 13 out of 68 studied hybrid apps (19%) have a difference in star ratings across platforms with at least a small effect size, which shows that, even though hybrid development tools advertise their "single codebase" selling point, some apps built using these tools still lack consistency in how users perceive them. There may be several reasons for this inconsistency. For example, users across platforms may have different expectations and preferences. Another possible reason is that one platform may have better alternatives for an app than the other platform. Finally, a possible reason is that hybrid development tools may generate higher quality apps for one platform over another.

Compared to native cross-platform apps, the consistency of star ratings across platforms of hybrid apps is much better. In our prior work on native cross-platform apps [17], we found that 14 out of 19 (74%) studied apps did not receive consistent star ratings across platforms.

The difference in the average of star ratings across platforms can be as large as two stars. We observe that the "Gartic" app receives an average rating of 4.6 in Android and 2.5 in iOS, the largest difference in the averages of star ratings in our study. Accordingly, the skewness of this app in Android indicates a strong right skew, while in iOS, the skewness shows a strong left skew. The skewness of star ratings of the "Gartic" app indicates that Android users rate the software quality high while iOS users are not satisfied with the software quality of

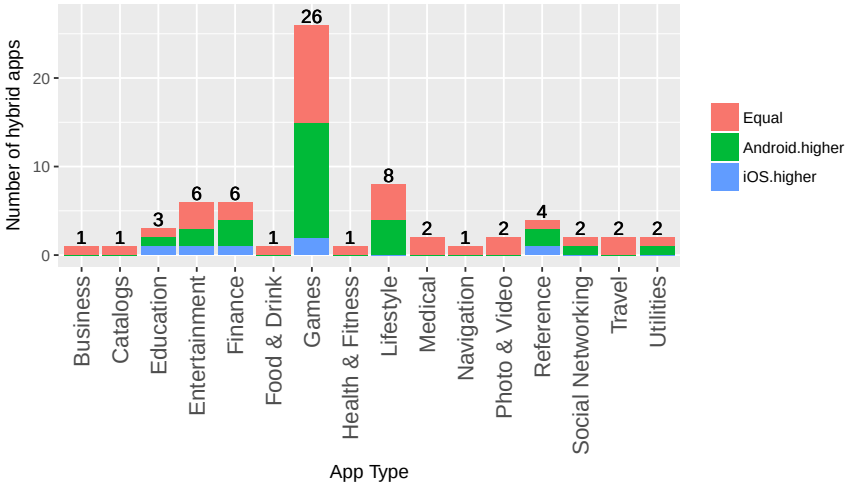

Fig. 2: The distribution over app types of apps that have no significant difference in ratings across platforms (*Equal*), apps that have a significantly higher rating on Android and apps that have a significantly higher rating in iOS.

the app. We manually examined the reviews and found that iOS users of the app complained a lot about the language of the app (e.g., *"Even though it let's you pick your language, the actual game is not in English, making it impossible to play. Fix this and it would be a good game, similar to pictionary."*) and a lack of players in the game (e.g., *"It is a interesting game, but it needs people to play with and there are too little players, I need to wait a long time and there still no one to show up...."*), while Android users do not complain about the above issues. A possible explanation is that the Android version is more popular than the iOS version (and hence there are more players), which is reflected by the number of collected reviews (625 in Android vs. 25 in iOS).

26 out of 68 (38%) studied hybrid apps are game apps and 24 of those 26 apps (92%) have significantly higher or the same star ratings in Android. Figure 2 shows that game apps are the most occurring type of app in our study. While games take up 25% of all apps in Apple's App Store [43] as well, the percentage of games among hybrid apps appears to be even larger. 24 out of those 26 hybrid game apps have higher or the same average ratings in Android. We could not identify a reason for the higher ratings in Android in the reviews nor in the release notes.

The percentage of hybrid finance apps is 4 times as large as the percentage of non-hybrid finance apps. 6 out of 68 (9%) studied hybrid apps are finance apps, compared to 2.3% across the whole app market [43]. Our explanation is that developers prefer a hybrid approach for their finance apps mainly because such apps are primarily an interface to an already existing web service. Hence, the transition to a web-based hybrid app is simple. We find that all 6 identified finance apps are backed by a website with a "log in" feature, which indicates that users have access to functionality that helps them to manage their accounts after

logging in. The popularity of building finance apps using hybrid development tools suggests that if developers wish to build cross-platform apps for their existing website, building a hybrid app can be a viable option. However, as the median of the average ratings of finance apps is 3.0 in Android and 2.7 in iOS, building a hybrid app does not necessarily mean that its users perceive the app of high quality.

Hybrid development tools do a fairly good job at achieving consistency in star ratings across platforms, in particular, in comparison to native cross-platform apps. However, approximately 1 out of 5 hybrid apps do not achieve such consistency.

4.2 How Consistent Are The Raised Topics In Reviews Rated Across Platforms?

Motivation: Knowing the popular topics in user reviews helps developers understand better what users care about. As discussed earlier, the star ratings play an important role in attracting potential users. Hence, we pay particular attention to topics that have an extreme (i.e., large and small) inconsistency in star ratings. For developers, knowing the topics that result in large inconsistency in star ratings could help them to identify important tasks related to the topics, and hence, prioritize their development efforts. Different from the first RQ in which we study the star ratings of hybrid apps that share the same codebase, we focus on whether such hybrid apps can achieve consistency across particular topics that are raised in user reviews across platforms in this RQ.
Approach: To find out what users discuss on cross-platform hybrid apps, we use Twitter-LDA [50] to extract discussion topics from all 1 to 5-star reviews for the identified cross-platform hybrid apps. Twitter-LDA has been widely used for topic modeling purposes in microblogs, such as Twitter. Compared to the original LDA by Blei [6], Twitter-LDA is able to identify more meaningful topics in short documents, such as tweets. In Twitter, each tweet can have a maximum of 140 characters. We find that on average, a review from an iOS user has 155 characters, and a review from an Android user has 65 characters. The number of characters of reviews from Android and iOS users suggests that using Twitter-LDA will yield better results than the original LDA.

Since Twitter-LDA is an unsupervised learning algorithm, it is difficult to analyze and compare topics that are returned by different runs of Twitter-LDA [44]. Combining user reviews of an app from the two studied platforms makes the comparison possible, as one common set of topics will be extracted from the reviews of the app on both platforms. Hence, we combine the user reviews of each hybrid app and run Twitter-LDA at the app level. For example, we run Twitter-LDA on the combined set of user reviews from Android and iOS for the app "Discover Mobile". We also filter out 13 out of 68 apps that have less than 50 reviews on one of the platforms in order to get enough reviews for each app. Hence, we run Twitter-LDA 55 times in total.

In order to ensure the meaningfulness of the results of the Twitter-LDA runs, we preprocess the user reviews through the following steps:

1. Change all words to lower case (e.g. *App* to *app*)
2. Remove punctuation
3. Remove English stop words (e.g. *the*)

Rank	App Name	Keywords of the topic	Average Rating		Ratio[1]	# of Reviews	
			Android	iOS		Android	iOS
1	Michigan Lottery Mobile	locat play updat verifi terribl scan iphon wifi anyth ticket	3.6	1.2	3.1	36	11
2	Bridge Base Online	bid partner onlin practic redeal bridgebas format horribl anoth superb	3.7	1.7	2.2	37	13
3	ElfYourself by OfficeMax	app updat crash open ipad fix email work send cool	4.6	2.3	2.0	26	46
4	Bridge Base Online	robot bid hand partner suit perfect live made diamond lead	4.1	2.1	1.9	22	13
5	ACTPhoto	tri crash time galleri everi wast close forc pointless recent	1.0	2	1.9	76	16
6	GameStop	love great gamer cool point stop awesom notif reward redeem	4.0	2.1	1.9	136	25
7	GameStop	card work store click time preorder show shop everi updat	3.8	2	1.9	71	40
8	GameStop	log sign password time everi account tri fix reset work	3.9	2.1	1.9	63	146
9	Schwan's Food Delivery	open crash load everi time anyth click close forc asu	4.4	2.3	1.9	51	12
10	GameStop	fix bluetooth time point star anymor search anyth spend constantli	3.9	2.1	1.9	52	47

[1]Ratio is calculated by the higher average star rating / the lower average star rating.

Table 4: Statistics for topics in Twitter-LDA that have the most significant inconsistency across platforms

4. Stem user reviews using the Porter stemmer [38] from the Python NLTK library (e.g. *meeting* to *meet*)

For user reviews that are assigned the same topic by Twitter-LDA, we examine the differences in star ratings across Android and iOS. To make the results more meaningful and unbiased, we only keep the topics to which at least 10 reviews are assigned on both platforms. We are interested in the topics that have large differences in terms of average star ratings across platforms. In our study, we only use the most important topic for each review (i.e., the topic with the highest topic probability).

Results: **For the same topic of user reviews of a hybrid app, the star ratings between Android and iOS may differ up to three times.** Table 4 shows the ten topics that have the largest difference in average ratings of an app across both platforms. The ratio in Table 4 is calculated as the higher average star rating for a particular platform divided by the lower average star rating for the same app on the other platform. The largest ratio in Table 4 is 3.1 for the Michigan Lottery Mobile app, which indicates that Android users give an average star rating that is 3 times higher than iOS users. For the topic that results in the largest ratio of average star ratings, the number of reviews of this topic is 11 in iOS and 36 in Android. We also observe that the Bridge Base Online and GameStop apps have two and four topics that appear in the top topics that have the most significant inconsistency in average star ratings, respectively. The above observations further suggest that the inconsistency in user reviews of hybrid apps across platforms may be severe and that developers of these apps should spend time to address this inconsistency.

We read the reviews of the topic for the app Michigan Lottery Mobile and we find that the iOS version has a worse user experience when validating lottery tickets than the Android version. For example, an iOS user gave a 1-star rating and complained that *"Unless I'm a moron, there is no tool to scan to check if a ticket is a winner. Puzzling since scanning functionality is there to enter ticket codes for the Player's Club. This seems a no brainier. Unless, again, it's there and I can't figure out how to use it."*. Another iOS user complained that *"9 out of 10 times it says that can't verify that I'm in Michigan I contacted tech support several times and no help"*. On the other hand, Android users praised the app and mentioned that *"It really works Very good ALWAYS."*. When we examined reviews of the Bridge Base Online app, we found that the app has a user interface problem on iOS after updating. For example, an iOS user gave a 2-star rating and complained: *"The app was nice, but it just updated and now the resolution is bad. What happened to the quality of the picture? I understand improving it for iPhone 5, but it should still work well on iPhone 4S. Please fix this."*

Understanding the cause of the above observations is difficult as the Android version of the apps could be intrinsically better than its iOS counterpart (for example, because of Android-specific features that were added over time). Another possibility is that the iOS operating system needs to better accommodate older apps on newer Apple phones. To understand the possible reasons of such great differences between two platforms even on the same topic, we examined the release notes of the Michigan Lottery Mobile and Bridge Base Online apps. We found that the ticket validation problem of the Michigan Lottery Mobile might only exist in the iOS version. A release note from Oct 23, 2013 states: "Barcode scanner bug

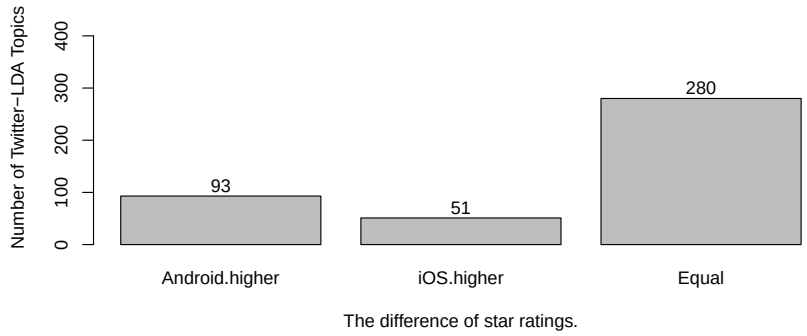

Fig. 3: Distribution of the difference between the average star ratings of Android and iOS for Twitter-LDA topics.

fixes for iOS 7." We could not find similar release notes for the Android platform, which implies that this bug might only exist on iOS 7.

We also identified that the user interface problem of the Bridge Base Online app is more notable on iOS than on Android. We identified several user interface improvements in the iOS version, such as "Improved screen layouts when playing or kibitzing in both pictures of cards and hand diagram mode" (version 3.85 on Aug 14, 2014), "Playing/kibitzing interface has been rewritten. The app should be much more usable on phones now." (version 3.70 on Jan 27, 2014), and "Graphics improvements for iPhone 4" (version 3.51 on Jan 07, 2013). We did not find such improvements for the Android version. In summary, we can conclude that even for hybrid apps, which were created using one codebase, inconsistencies in user reviews exist and can be considerable, as demonstrated by the Michigan Lottery Mobile and the Bridge Base Online apps.

144 out of 424 (34%) topics are not equally rated across platforms. In total, Twitter-LDA identifies 424 topics from the reviews of the 55 studied hybrid apps after removing the topics that have less than 10 reviews. For each Twitter-LDA topic, we first use the Mann-Whitney U test (p-value < 0.05) to decide whether the star ratings are significantly different across platforms. Then, we calculate the ratio of average star ratings using the average star rating of reviews in Android divided by the average rating in iOS to decide on which platform the ratings are higher.

Figure 3 shows the number of topics that are rated higher in Android, higher in iOS or equally rated across platforms. We find that for 93 Twitter-LDA topics, the average star ratings of the reviews under these topics are significantly higher in Android. In contrast, for 51 Twitter-LDA topics, the average star ratings of the reviews matching these topics are significantly higher in iOS. 280 Twitter-LDA topics receive equivalent average star ratings across platforms. Hence, 144 out of 424 (34%) topics are not rated equally across platforms.

The fact that almost twice as many Twitter-LDA topics receive significantly higher star ratings in Android reviews may suggest that Android users are more generous in giving high star ratings, or less picky about issues that arise. Other possible explanations are better support of a platform for hybrid apps, or different expectations of users across platforms. More importantly, our findings indicate that apps that were created by hybrid development tools do not often receive consistent star ratings even for the same topics that users are discussing across platforms. However, as argued previously, it is difficult to understand the root cause of the above observation. Therefore, developers should not solely rely on hybrid development tools as a method to ensure consistency in user reviews.

Hybrid development tools do not guarantee consistency in user reviews across platforms: 34% of the discussed topics in reviews do not receive consistent ratings across platforms.

5 Implications of Our Findings

In this section, we discuss the implications of our findings.

Developers that wish to offer apps that are consistent across platforms should use a hybrid development framework. Hybrid development frameworks produce apps that achieve a better consistency in terms of star ratings than native cross-platform apps. In our previous study on native cross-platform apps [17], we found that 14 out of 19 (74%) studied apps did not receive consistent star ratings across platforms. As shown in Section 4.1, the percentage drops to 47% when the study subjects change from native cross-platform apps to hybrid apps. In particular, when considering only significant differences with at least a small effect size, only 19% of the hybrid apps receive inconsistent ratings across platforms. Our findings suggest that hybrid development tools do a fairly good job at delivering apps that are rated consistently across platforms. However, it is important to keep in mind that these tools do not necessarily guarantee such consistency. In order to achieve such consistency, developers need to carefully track issues that are raised in reviews across all supported platforms and act accordingly.

One thing that developers should keep in mind is that hybrid apps appear to be better received on Android. 75% of the apps that have a significant difference in their star ratings across platforms have higher average star ratings in Android. This percentage is similar to what we observed for native cross-platform apps (65-80%) in our prior work [17]. As discussed in our prior work, users across different platforms complain about different issues [17]. One possible explanation is that users across platforms have different expectations. For example, as iOS users tend to be more interested in technology than Android users [5], they may expect an app to quickly support new features of the latest iPhone model. Hybrid development tools may not support such features, as these features may not be available on all supported platforms.

Developers of hybrid apps might still require platform-specific coding to address platform differences. Unfortunately, hybrid development tools do not generate flawless apps on all platforms. Section 4.2 shows that users rate the same topic inconsistently across platforms. In particular, our study shows that for several apps, release notes on iOS show that several improvements or bug fixes were made while no such improvements were necessary. One possible explanation

is that release notes across platforms are written by different developers, which do not provide the same level of information their release notes. However, as the differences in ratings across platforms are substantial, a more likely explanation is that hybrid apps may deliver apps that contain inconsistent bugs on one of the supported platforms. Hence, developers need to thoroughly test their apps across platforms, and may still need to write platform-specific code to address platform-specific bugs. In addition, they need to track reviews across all supported platforms to quickly identify platform-specific issues.

Developers should prioritize their bug fixing efforts differently across platforms, as the same issue has a different impact on the rating across platforms. Section 4.2 shows that users rate the same issue differently across platforms. Hence, hybrid app developers must identify the most important issues for their users across platforms, and prioritize bug fixing efforts accordingly. As a result, the same issue may be more urgent to fix on one platform than on another.

6 Threats to validity

6.1 External validity

The findings of our RQ1 highlight the risk of analyzing the consistency in star ratings at the platform level since such analysis reveals limited information for developers to improve their own hybrid apps. Instead, developers should focus on comparing their hybrid apps on an app versus app basis. While our specific findings might not generalize, our findings do highlight the existence of differences across platforms for the same hybrid app. However, all our proposed techniques and our methodologies are general and can be used for any app.

We use an existing dataset [3] to select free-to-download hybrid apps. The majority of the hybrid apps in this dataset are non-popular "zombie apps". Therefore, we focus our study on hybrid apps that have at least 100,000 downloads in Android (note that the App Store does not provide data about download numbers) and 20 reviews on each studied platform. As a result, our findings apply only to free-to-download popular hybrid apps. Future research is necessary to study whether our results are valid for paid or non-popular hybrid apps.

6.2 Internal validity

In this paper, we configured Twitter-LDA with the same setting for all cross-platform hybrid apps. As mentioned earlier, we set the number of topics to 10 and the number of iterations to 1,000. In addition, we selected the most important topic for each review. These choices change the granularity of the topics only. As we are interested in the comparison of the topics across platforms, and not in the exact topics, these choices did not affect the outcome of our study. In addition, we only examined topics that match at least 10 topics on each platform. Using a different number of topics and number of iterations may yield different results. However, there is no rule of thumb that tells us the optimal parameters of Twitter-LDA. We observed that when setting the number of topics to 10, the results from running Twitter-LDA on several cross-platform hybrid apps have better interpretability

compared to other settings such as 5 or 15. Therefore, we used the same parameters for all executions of Twitter-LDA.

In this paper, we assumed that the inconsistency across platforms was caused by the hybrid development tool. However, there may be many reasons for the inconsistency in star ratings and reviews. For example, in our prior work on native cross-platform apps [17], we observed that there exist differences between users across platforms. In addition, native cross-platform apps are often built and maintained by different development teams across platforms. However, hybrid apps are a special kind of cross-platform app. The developers of hybrid apps specifically choose to not have different code bases across platforms (i.e., because they use a hybrid app development tool). There is no reason for hybrid app developers to have inconsistency that is not caused by the internal web browser across platforms, or by the code that is generated by the hybrid app tool. In addition, we observed a clear improvement in consistency of star ratings and reviews across platforms when comparing hybrid apps to natively-built cross-platform apps. However, more research is necessary to study which other factors may influence the inconsistency in users' impression across platforms.

In this paper, we focused on absolute metrics to study the consistency of star ratings across stores. An alternative would have been to use relative metrics, such as the relative ranking in top lists across stores. However, we cannot compare relative metrics across stores, due to differences between the stores. For example, imagine App A has ranking 5 in the Google Play store, and ranking 4 in the App store. These rankings are inconsistent, but their comparison is not that meaningful because the distribution over categories across platforms is different. On one platform, there may be more apps that offer the same functionality as app A than on the other platform. Hence, the competition for an app may be higher on one platform, which could have its effect on the rankings. Therefore, unfortunately relative metrics do not give an accurate view of consistency, and we did not include them in the paper.

7 Conclusions

Analyzing the star ratings and reviews of hybrid apps, i.e., mobile apps that are developed using hybrid development tools that are available on multiple platforms, provides app developers a unique insight of how app users across different platforms perceive the software quality of their apps. The majority of prior work on mobile apps is done from a developer's perspective or limits the app selection to one app store.

In this paper, we study the software quality of hybrid apps from a user's perspective. We study 68 popular free-to-download hybrid apps and find that they do a better job at receiving consistent star ratings across platforms in comparison to cross-platform native apps (i.e., cross-platform apps that were built by native development kits). However, when we look at the actual contents of user reviews using topic modeling (Twitter-LDA), we find that for the same topic of user reviews, users on different platforms can react differently in terms of star ratings. The main contributions of our study are:

1. We demonstrate an approach that combines star rating and user reviews to evaluate the consistency of hybrid apps across platforms.

2. We show that apps built by hybrid development tools still lack consistency that needs to be addressed by developers. In particular, developers need to closely track the star ratings and reviews that their apps receive across platforms in order to identify and address platform-specific issues.

Even though building mobile apps using hybrid development tools can save time and reduce the development cost of an app, the consistency in star ratings and user reviews is still not 100% assured by simply adopting hybrid development tools. However, the consistency is better than that of natively-built cross-platform apps. Hence, hybrid development tools appear to offer a good starting point for achieving consistency in users' impression of an app.

An interesting question is whether hybrid development tools help developers to build high quality apps, aside from consistent apps. Given the fact that only a small portion of the most popular apps are hybrid ones, there appears to be a relation between the popularity of an app and the usage of a hybrid development tool to build the app. This observation is supported by the relatively low ratings that are given to the studied hybrid apps. Future studies should be done to deeper investigate the quality of hybrid apps.

8 REFERENCES

1. Adobe (2017) Phonegap. `https://phonegap.com/`, (last visited: Oct 3, 2017)
2. Akdeniz (2014) Google Play Crawler JAVA API. `https://github.com/Akdeniz/google-play-crawler`, (last visited: Jan 25, 2017)
3. Ali M, Mesbah A (2016) Mining and characterizing hybrid apps. In: Proceedings of the International Workshop on App Market Analytics (WAMA), ACM, pp 50–56
4. Apple (2008) RSS feed provided by Apple for the app "Facebook". `https://itunes.apple.com/us/rss/customerreviews/id=284882215/page=1/json`, (last visited: Jan 25, 2017)
5. Benenson Z, Gassmann F, Reinfelder L (2013) Android and iOS users' differences concerning security and privacy. In: Extended Abstracts on Human Factors in Computing Systems (CHI), pp 817–822
6. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
7. Chen N, Lin J, Hoi SCH, Xiao X, Zhang B (2014) Ar-miner: Mining informative reviews for developers from mobile app marketplace. In: Proceedings of the 36th International Conference on Software Engineering (ICSE), ACM, New York, NY, USA, pp 767–778
8. Dalmasso I, Datta SK, Bonnet C, Nikaein N (2013) Survey, comparison and evaluation of cross platform mobile application development tools. In: 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC), pp 323–328
9. Di Sorbo A, Panichella S, Alexandru CV, Shimagaki J, Visaggio CA, Canfora G, Gall HC (2016) What would users change in my app? Summarizing app reviews for recommending software changes. In: Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE), ACM, pp 499–510
10. Fu B, Lin J, Li L, Faloutsos C, Hong J, Sadeh N (2013) 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 (KDD), ACM, New York, NY, USA, pp 1276–1284
11. Graphpad Software (2015) Interpreting results: Skewness and kurtosis. `http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_skewness_and_kurtosis.htm`, (last visited: Jan 30, 2016)
12. Gu X, Kim S (2015) What parts of your apps are loved by users? In: 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp 760–770

13. Guzman E, Maalej W (2014) How do users like this feature? [a] fine grained sentiment analysis of app reviews. In: 22nd International Requirements Engineering Conference (RE), IEEE, pp 153–162
14. Harman M, Jia Y, Zhang Y (2012) App store mining and analysis: MSR for app stores. In: 9th Working Conference on Mining Software Repositories (MSR), IEEE, pp 108–111
15. Hassan S, Shang W, Hassan AE (2017) An empirical study of emergency updates for top Android mobile apps. Empirical Software Engineering (EMSE) 22(1):505–546
16. Heitkötter H, Hanschke S, Majchrzak TA (2013) Evaluating Cross-Platform Development Approaches for Mobile Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 120–138
17. Hu H, Bezemer CP, Hassan AE (2016) Studying the consistency of star ratings and the complaints in 1 & 2-star user reviews for top free cross-platform Android and iOS apps. `https://peerj.com/preprints/2589/`
18. Joanes DN, Gill CA (1998) Comparing measures of sample skewness and kurtosis. Journal of the Royal Statistical Society Series D (The Statistician) 47(1):183–189
19. Joorabchi M, Mesbah A, Kruchten P (2013) Real challenges in mobile app development. In: International Symposium on Empirical Software Engineering and Measurement (ESEM), IEEE/ACM, pp 15–24
20. Joorabchi ME, Ali M, Mesbah A (2015) Detecting inconsistencies in multi-platform mobile apps. In: IEEE 26th International Symposium on Software Reliability Engineering (ISSRE), pp 450–460
21. Khalid H, Shihab E, Nagappan M, Hassan AE (2015) What do mobile app users complain about? IEEE Software 32(3):70–77
22. Long JD, Feng D, Cliff N (2003) Ordinal Analysis of Behavioral Data, John Wiley & Sons, Inc.
23. Man Y, Gao C, Lyu MR, Jiang J (2016) Experience report: Understanding cross-platform app issues from user reviews. In: IEEE 27th International Symposium on Software Reliability Engineering (ISSRE), pp 138–149
24. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 18(1):50–60
25. Martin W, Harman M, Jia Y, Sarro F, Zhang Y (2015) The app sampling problem for app store mining. In: Proceedings of the 12th Working Conference on Mining Software Repositories (MSR), IEEE Press, pp 123–133
26. Martin W, Sarro F, Harman M (2016) Causal impact analysis for app releases in Google Play. In: Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE), ACM, New York, NY, USA, pp 435–446
27. Martin W, Sarro F, Jia Y, Zhang Y, Harman M (2016) A survey of app store analysis for software engineering. IEEE Transactions on Software Engineering (TSE) PP(99):1–32
28. Microsoft (2017) Xamarin: Mobile app development and app creation software. `https://www.xamarin.com/`, (last visited: Oct 3, 2017)
29. NIST/SEMATECH (2012) e-handbook of statistical methods: Measures of skewness and kurtosis. `http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm`, (last visited: Oct 3, 2017)
30. Noei E, Syer MD, Zou Y, Hassan AE, Keivanloo I (2017) A study of the relation of mobile device attributes with the user-perceived quality of Android apps. Empirical Software Engineering (EMSE) pp 1–29
31. Ohrt J, Turau V (2012) Cross-platform development tools for smartphone applications. Computer 45(9):72–79
32. Pagano D, Maalej W (2013) User feedback in the appstore: An empirical study. In: 21st International Requirements Engineering Conference (RE), IEEE, pp 125–134
33. Palmieri M, Singh I, Cicchetti A (2012) Comparison of cross-platform mobile development tools. In: Intelligence in Next Generation Networks (ICIN), 2012 16th International Conference on, pp 179–186
34. Palomba F, Linares-Vsquez M, Bavota G, Oliveto R, Penta MD, Poshyvanyk D, Lucia AD (2015) User reviews matter! tracking crowdsourced reviews to support evolution of

successful apps. In: 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp 291–300

35. Palomba F, Salza P, Ciurumelea A, Panichella S, Gall H, Ferrucci F, De Lucia A (2017) Recommending and localizing change requests for mobile apps based on user reviews. In: Proceedings of the 39th International Conference on Software Engineering (ICSE), IEEE Press, Piscataway, NJ, USA, pp 106–117
36. Panichella S, Sorbo AD, Guzman E, Visaggio CA, Canfora G, Gall HC (2015) How can I improve my app? Classifying user reviews for software maintenance and evolution. In: International Conference on Software Maintenance and Evolution (ICSME), IEEE, pp 281–290
37. Pettey C, Rob van der M (2012) Gartner says free apps will account for nearly 90 percent of total mobile app store downloads in 2012. http://www.gartner.com/newsroom/id/2153215, (last visited: Jan 28, 2016)
38. Porter MF (1997) Readings in information retrieval. Morgan Kaufmann Publishers Inc., chap An Algorithm for Suffix Stripping, pp 313–316
39. Poschenrieder M (2015) 77% will not download a retail app rated lower than 3 stars. https://blog.testmunk.com/77-will-not-download-a-retail-app-rated-lower-than-3-stars/, (last visited: Jan 28, 2016)
40. Ramon L, Ryan R, Kathy N (2015) Smartphone OS market share, 2015 q2. http://www.idc.com/prodserv/smartphone-os-market-share.jsp, (last visited: Jan 25, 2017)
41. Romano J, Kromrey JD, Coraggio J, Skowronek J, Devine L (2006) Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices. In: Annual meeting of the Southern Association for Institutional Research
42. Smutn P (2012) Mobile development tools and cross-platform solutions. In: Carpathian Control Conference (ICCC), 2012 13th International, pp 653–656
43. Statista (2017) Most popular Apple App Store categories in July 2017, by share of available apps. https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/
44. Thomas SW, Adams B, Hassan AE, Blostein D (2011) Modeling the evolution of topics in source code histories. In: Proceedings of the 8th Working Conference on Mining Software Repositories (MSR), ACM, pp 173–182
45. Tian Y, Nagappan M, Lo D, Hassan AE (2015) What are the characteristics of high-rated apps? A case study on free Android applications. In: IEEE International Conference on Software Maintenance and Evolution (ICSME), pp 301–310
46. Vashistha C (2015) Native vs hybrid mobile app: 5 ways to choose right plarform. URL http://www.ipragmatech.com/native-hybrid-mobile-app-right-platform/
47. Viennot N, Garcia E, Nieh J (2014) A measurement study of Google Play. SIGMETRICS Perform Eval Rev 42(1):221–233
48. Villarroel L, Bavota G, Russo B, Oliveto R, Di Penta M (2016) Release planning of mobile apps based on user reviews. In: Proceedings of the 38th International Conference on Software Engineering (ICSE), ACM, New York, NY, USA, pp 14–24
49. Vu PM, Nguyen TT, Pham HV, Nguyen TT (2015) Mining user opinions in mobile app reviews: A keyword-based approach. In: 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp 749–759
50. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing Twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR), Springer-Verlag, pp 338–349