Understanding the Challenges of Mobile Phone Usage Data

Karen Church
Yahoo Labs
kchurch@yahoo-inc.com

Denzil Ferreira
University of Oulu
denzil.ferreira@ee.oulu.fi

Nikola Banovic
Carnegie Mellon University
nbanovic@cs.cmu.edu

Kent Lyons
Yahoo Labs
dr.kent.lyons@gmail.com

ABSTRACT
Driven by curiosity and our own three diverse smartphone application usage datasets, we sought to unpack the nuances of mobile device use by revisiting two recent Mobile HCI studies [1, 17]. Our goal was to add to our broader understanding of smartphone usage by investigating if differences in mobile device usage occurred not only across our three datasets, but also in relation to prior work. We found differences in the top-10 apps in each dataset, in the durations and types of interactions as well as in micro-usage patterns. However, it proved very challenging to attribute such differences to a specific factor or set of factors: was it the time frame in which the studies were executed? The recruitment procedure? The experimental method? Using our somewhat troubled analysis, we discuss the challenges and issues of conducting mobile research of this nature and reflect on caveats related to the replicability and generalizability of such work.

Author Keywords
Mobile usage; Smartphone usage; Device usage; Evaluation; Methodology; User Studies; Mobile HCI; Micro-usage; Replication; Generalizability

General Terms
Human Factors

ACM Classification Keywords
H.5.2 Information interfaces and presentation: Miscellaneous

INTRODUCTION

Over the past decade, we have seen a steady increase in the number of real-world user studies of mobile device use. Recent research focuses on the nuances of smartphone usage and sheds light on the apps that people engage with, the duration of their usage and more recently what happens during very short, bursty device interactions [28]. To understand such behaviors a range of methodologies have been adopted (e.g., mobile diaries, Experience Sampling Method (ESM), Day Reconstruction Method (DRM), just to name a few) and in some cases, novel methods have emerged. One such method is the use of mobile loggers, deployed as applications in both smaller scale user studies as well as in large-scale app market deployments.

Despite the flurry of research in this space, most researchers will agree that studying mobile device usage is still very challenging. We have the traditional issues associated with conducting HCI studies (e.g., recruiting, incentives), confounded with studying the highly mobile (and therefore contextually varied) technology that also evolves rapidly. In smaller scale mobile studies, researchers often combine logging mobile usage with qualitative insights and follow a set of key steps to accommodate this mixed-method approach. This includes recruiting participants - in most cases locally - building custom app logging tools, handing out or configuring mobile devices, as well as offering incentives and rewards. The effort involved in such studies is often substantial due to the recruitment process that needs to take place as well as the compensation of the participants, especially if we aim at a longitudinal deployment.

Alternatively, the increasing popularity of application stores has allowed researchers to use such distribution mechanisms to recruit thousands of participants and conduct large-scale studies in naturalistic settings. However, there are potential side-effects and biases that such an approach may introduce and thus researchers following such paths need to consider a number of factors to ensure the success of their study. For example, how polished should an application (i.e., logger) be in order to attract participants; how does device hardware diversity affect data sampling; how biased is the participant pool given their interest in the application's functionality in the first place; and what prevents participants from removing the application for longitudinal analysis? [19].

Regardless of the challenges, these studies contribute to a broader understanding of mobile phone usage behaviors as well as providing insights about future areas of exploration in Mobile HCI. In this paper, we were interested in uncovering additional user practices in the mobile space using three different, but similar app logging datasets: two gathered through small-scale studies; and one larger-scale app market deployment. Our studies cover different time frames

MobileHCI '15, August 25 - 28, 2015, Copenhagen, Denmark
2015 ACM. ISBN 978-1-4503-3652-9/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2785830.2785891

(between 2013-2014), locations (in the US as well as round the world) and user populations (teenagers, adults, iOS users, Android users). Our goal was to revisit prior work on short mobile device interactions [1] and short, bursty app usage, so called *micro-usage* [17], and to determine what (if any) differences emerge both between the datasets and also in relation to prior related work. While we did find several interesting differences, in retrospect it proved very challenging to attribute these differences to a specific factor or set of factors: Is it the differences in setting, technology, applications, timing or are the differences rooted more in the user populations sampled?

Thus what has started as a replication study, quickly turned to a deeper reflection on the methods and challenges surrounding how we conduct mobile research ourselves and as a community. Our core contribution then lies in discussing what it really means to do research in this entropic and volatile setting. To start, we review several pieces of research that use similar methodological approaches to ours and summarize some of the practices the community is adopting. We then present our study by analyzing three app usage datasets that measured similar data across three different user populations and highlight how a range of issues emerge in this type of analysis. Lastly, we discuss what our findings mean more broadly for the research in our community in terms of generalizability and replicability. Our major insight is that while data of this nature provides great value in understanding how people use their mobile devices and informing the design of novel interactions that improve mobile device interactions, these insights comes with boundaries and caveats which all of us must be aware of.

CHARACTERIZING MOBILE DEVICE USAGE STUDIES

Mobile phones and their built-in sensors have increasingly transitioned research to users' pockets instead of a laboratory. Given the challenging nature of mobile environments, some researchers have focused their efforts on lowering the burden of conducting qualitative mobile studies "in-the-wild," by introducing novel methods to collect such data from participants and capture a glimpse into their daily lives, including: voicemail diaries [2, 29], snippet-based diaries [4, 9], refined and context-aware experience sampling methods [8, 9, 10, 21, 24, 42] and well as novel video collection methods [5]. Other researchers have focused on building sensing tools and techniques [7, 22, 33, 34, 36], and even complete mobile sensing frameworks (e.g., Funf[1] and AWARE[2]) for continuously and unobtrusively gathering mobile usage and sensor data. Such frameworks have provided the community with the capabilities to study actual mobile device usage in a streamlined manner at unprecedented scale (e.g., [3, 15, 44]) or over long periods of time (e.g., [12, 35, 37]).

Researchers quickly started deploying their studies to different populations to explore and understand specific aspects of mobile device usage. For example, studies focused on negative aspects of mobile device use among specific demographics, with teenagers and college students attracting much attention [26, 38]. The Mobile HCI research community took particular interest in studies focusing on the intricate details of application and device use. Such studies focused on how and where users interact with applications on their mobile devices (e.g., [5]), and also for how long (e.g., [17]). These studies have informed the design of new interactions with mobile devices by utilizing known patterns of mobile device use.

Knowledge about users' application use patterns has also led to the development of smart home screens and launchers that help improve device battery life [14, 16] or reduce application invocation time by predicting the user's next application (e.g., [25, 38, 39, 43, 45, 46]). Similarly, the nuances of mobile device use, for example, understanding smartphone notifications [32], has informed the design of systems that reduce unwanted interruptions [30, 31]. Understanding how people interact with their mobile device lock screens and home screens has also led to design of novel interactive lock screens that power crowdsourcing efforts [40, 41], streamline short interactions with mobile devices [1] and help people learn new languages [11]. Studies to evaluate these new interventions often deploy the intervention to the users' mobile devices and track the user's interaction with the intervention in the field; thus keeping the cycle of collecting large amounts of usage data going.

The result of such studies is that researchers are now faced with large amounts of mobile device usage data, from various populations, with each study focused on understanding different aspects of mobile device usage. In addition to all the challenges of analyzing such large amounts of data, researchers are doing their best to follow the rigor of the scientific method in presenting their findings. However, the complexity of this research field can make this a difficult problem for the Mobile HCI community. Therefore, it is no surprise to find that our community has already identified, discussed, and made an attempt to provide best practices to deal with some of the challenges of conducting mobile device usage experiments in particular in large-scale, in-the-wild deployments [19, 23, 27, 37, 44]. Regardless of these challenges, an expectation that the findings of mobile device usage studies are generalizable and the results replicable remains for the large part.

To grasp the complexity of understanding mobile device usage, it is important to consider the many different aspects of past studies in this field. These aspects include: 1) the current state of mobile device research, industry and market at the time of the study, 2) the research goals of the researchers given that context, and 3) the study design used to collect data to meet the research goals. By carefully considering each of these aspects, researchers are able to apply existing community domain knowledge to their particular research or design goals. Understanding the various aspects of prior studies also helps them reproduce, and hopefully replicate prior studies using new datasets.

Table 1 provides a visual review of related studies of mobile app usage. Our review of existing research was driven by our initial research goal of gaining a more generalizable understanding of mobile device use across different user populations. We selected the studies to review on the basis of their similarity and comparability to our study goal. Each study

[1]See www.funf.org

[2]See www.awareframework.com

Research goals	Reference	Year	Number of participants	Duration	Deployment P	Deployment S	Recruitment L	Recruitment D	Incentives Yes	Incentives No	Method I	Method E	Data collected	User input
Streamline short mobile device use	Banovic *et al.*, 2014	2014	40	4 weeks	•		•		•		•		• • • • • • •	interview
Understanding of mobile phone use in everyday life	Brown *et al.*, 2014	2014	15	4 weeks		•	•		•			•	• • •	interview, diary
Understanding of application micro-usage and context in everyday life	Ferreira *et al.*, 2014	2014	36	5 weeks		•	•		•			•	• • • • • • •	interview, ESM
Understanding of mobile phone addiction by college students	Lee *et al.*, 2014	2014	95	4 months		•	•		•			•	• • • • • •	interview, survey
Understanding of mobile phone notifications	Pielot *et al.*, 2014	2014	15	1 week		•		•	•			•	• • • •	diary
Methodology to measure real-world, longitudinal smartphone use	Wagner *et al.*, 2014	2014	12500	2 years	•			•		•		•	• • • •	
Mining frequent patterns on the phone	Srinivasan *et al.*, 2014	2014	106	6 months	•			•		•		•	• • • • • • • •	survey
Improve battery management	Ferreira *et al.*, 2013	2013	12; 22	4 weeks; 3 weeks		•	•		•		•		• • • •	interview
Application recommendations	Zhang *et al.*, 2013	2013	7	6 weeks		•	•			•	•		• • •	
Understanding of successful experiments on appstores	Henze *et al.*, 2012	2012	8; 670; 4197; 5103; 6907	8.5; 6; 6; 5.5; 2 months	•			•		•		•	• • •	review, feedback
Understanding of smartphone use habits	Oulasvirta *et al.*, 2012	2012	136;15;12	3 months; 4 months; 2 weeks		•	•	•		•	•	•	• • • • •	survey; interview; DRM, interview
Launcher that predicts the next application to be used	Yan *et al.*, 2012	2012	Shepard et al., 2011 + 3	1 month		•	•			•	•		• • •	
Understanding of mobile charging behavior	Ferreira *et al.*, 2012	2012	Ferreira et al., 2011	4 weeks	•			•		•	•		• • •	review, feedback
Launcher that predicts the next application to be used	Shin *et al.*, 2012	2012	23	1 month	•			•		•	•		• • • • • •	
Understanding of mobile application usage behavior	Böhmer *et al.*, 2011	2011	4125	167 days	•			•		•		•	• • •	
Understanding of applications and context in everyday life	Do *et al.*, 2011	2011	77	9 months		•	•			•		•	• • • •	
Methodology to measure real-world, longitudinal smartphone use	Shepard *et al.*, 2011	2011	25	1 year		•	•		•			•	• • • • • •	interview
Understanding battery management	Ferreira *et al.*, 2011	2011	4437	4 weeks	•			•		•		•	• • • •	
Understanding of mobile phone use impact in battery life	Falaki *et al.*, 2010	2010	255	28 weeks	•			•		•		•	• • • •	
Understanding worldwide trials of mobile systems	McMillan *et al.*, 2010	2010	8676	5 months	•			•		•		•	• • •	feedback, social network
Using machine-learning to predict users' mobile operations	Kamisaka *et al.*, 2009	2009	19	8 months		•	•			•		•	• • • •	
Understanding of mobile phone use in teenage users	Rahmati & Zhong, 2013	2007	14	4 months		•	•		•			•	• • • •	focus group, interview

Deployment: **P** - Public **S** - Selected
Recruitment: **L** - Local **D** - Distributed
Method: **I** - Intervention **E** - Experiment

ESM: Experience Sampling Method
DRM: Day Reconstruction Method

Time, Communication, Mobility, Battery, Screen, Notifications, Multimedia, Applications, Locations, Touch

Table 1. Summary table to help visualize the various practices employed by related studies of mobile device usage. The table shows: (1) Number of participants: recruited participants, separated by semi-colon if paper reports more than one study; (2) Duration: length of deployment, per study; (3) Deployment: if Public (i.e., crowdsourced, social networks, appstores) or Selected (i.e., recruited); (4) Recruitment: if Local (i.e., university, city) or Distributed (i.e., spread across the globe); (5) Incentives: participation rewards; (6) Method: Intervention or Experiment, i.e. (Naturalistic observation conducted in the field); (7) Data collected: sensor data collected (i.e., quantitative); and (8) User input: interactions with the participants (i.e., qualitative / subjective input).

we reviewed uses some form of mobile app logger and was conducted in-situ in the last 8 years. Overall this synthesis of related studies helped us identify the various aspects of mobile device use studies and enables us to compare and contrast the different studies.

The table is sorted by the reported deployment year (when available, or publication year otherwise), and includes a range of details including: a reference and year, the research goals, the study duration, the study method, as well as details related to participant recruitment. As expected, we found a wide spectrum of research goals across the studies. Over the years, researchers tend to re-visit similar research goals, perhaps to assess if users' behavior has changed in response to technological advances over time. With some exceptions, we also find longer deployments, varying from months to years. Lastly, we also observe that the majority of our literature review methodology are quasi-experiments or naturalistic observations conducting in the field (i.e., unobtrusively collecting data) rather than interventions (i.e., introduction of some novel application or interaction technique) — 14 vs 8 studies, respectively. We'll return to this review table throughout our discussion of the challenges and implications of conducting research of this nature.

REVISITING MOBILE DEVICE USAGE: OUR DATASETS

Our original goal was to replicate recent research on smartphone usage using three new datasets that all logged application use in a similar way. We wanted to systematically explore the similarities and differences between application usage characteristics to see what we could discover regarding mobile phone usage. Our premise was that, by combining these three diverse datasets and reflecting upon these different populations, we would obtain a better sense of overall mobile behaviors. To do so we examined the application usage data from three studies of smartphone usage we conducted between 2013 and 2014:

- A longitudinal application logging study conducted over a 6-month period (January-June, 2014) with 87 participants. We refer to this as the *Securacy* study.

- A smaller-scale 2-week mixed-method study of 20 active mobile search users conducted between June-July 2014. We refer to this as the *MSearch* study.

- A 2-week mixed method study of 14 teenagers conducted between August-September 2013. We refer to this as the *Teen* study.

Securacy Study

In the beginning of the year 2014, we conducted an in-the-wild 6-month long deployment with 87[3] anonymous app store users. We attempted to improve the end-user management of privacy and security concerns on mobile phones while they use applications. Note additional details from the full Securacy study is reported in [18]

Demographics

We built and deployed the Securacy application on the Google's Play Store, and for recruitment we advertised it on Facebook and Twitter. We offered no additional compensation beyond use of the Securacy app itself. The majority of our anonymous participants were from Europe (41%) and North America (53%). Our participant sample might not be representative of the general population, as it is likely that our participants had an accentuated interest in their mobile applications' network usage, privacy and security.

Data collection

We used AWARE [20] to build our logging application and collect our data. AWARE is an accessibility service, thus is constantly active on participants' devices. We unobtrusively collected time stamped usage data including foreground and background applications, screen on/off events, device lock/unlock events, network connections and location information. Qualitative data regarding users' perceptions of applications' security were collected via ratings and a mobile questionnaire assessing users' rational on rating score/scale.

App Usage Data

During a 6-month long deployment (January-June, 2014), 87 anonymous users logged a total of 152,275 application launches (M=1009.1 per participant). We logged a total of 2,761 unique applications across the 87 participants.

MSearch Study

In the summer of 2014 we conducted a 2-week mixed-method study with 20 Android users aged between 18 to 59 to explore the triggers, actions and contextual factors influencing mobile search. Our specific focus was on understanding how mobile search and general app usage interacts/relates. Note that full results from a subset of users (18/20) from this study can be found in [6].

Demographics

We recruited 10 males and 10 females, all of whom used an Android smartphone. Participants had an average age of 34 (SD=10.8). Participants were recruited from 9 different cities around the greater San Francisco Bay Area. Participants had a diverse set of occupations including students, administrative assistants, social workers, managers, chemists, homemakers and construction workers. Their education levels ranged from high school to college degree. All participants were active users of mobile search engines like Google and Yahoo.

Data Collection

App usage data was collected via a custom-built Android logging app that was installed on participants phones during an in-person interview. The logging app ran as an Android service in the background of the participants phone and kept track of all their mobile app usage. This tool collected time stamped usage data, specifically: which app was currently active, the time it was launched, and how long that app stayed active in the foreground. It also tracked events like display on and off as well as when the user accessed the home screen of the phone. Qualitative insights were collected via both in-person interviews (conducted at the start and end of the 2-week study period) as well as daily online diaries.

App Usage Data

App usage usage logs were gathered between 23rd June and 13th July 2014. The loggers ran for an average of 16 days on each participant's mobile phone. We collected 57,858 app opens/launches (M=2892.9 per participant) and we logged a total of 419 unique apps across the 20 participants.

Teen Study

In the summer of 2013 we conducted a 2-week mixed-method study with 14 teens aged between 13 to 18 to explore how they used their smartphones their daily life.

Demographics

We recruited 5 males and 9 females, half of whom used an iPhone and half of whom used an Android smartphone. Half of the participants were between 13-15 years old, while half were 16-18 years old. Participants were recruited from 11 different cities around the greater San Francisco Bay Area to represent various ethnic and socio-economic backgrounds. To mimic the greater US population, 2/3 of participants were selected such that neither parent had a university degree. All participants spoke English fluently although several spoke other languages at home. Parental consent was obtained for all participants under 18.

Data Collection

App usage data was again collected via custom-built logging applications, one for Android and one for iOS. The logging app was installed on participants phones during an in-person interview at the start of the study. This logging app ran in the background of the participants phone and kept track of all their mobile app usage. This tool again collected time stamped usage data, specifically: which app was currently active, the time it was launched, and how long that app stayed active in the foreground. It also tracked events like display on and off as well as when the user accessed the home screen of the phone. In-person interviews were conducted both at start and end of the 2-week study period and participants left daily voice mail diaries to tell us more about the details of their day-to-day app usage patterns.

[3]Note we filtered the Securacy dataset so that we only include users that had at least 1 day of mobile application use data. This resulted in 87 users in the Securacy study out of an initial pool of 218 users. We did this filtering to diminish the potential effects of outliers in our later analysis (e.g., once off device uses during the time the participants are considering whether to keep the study software or participants who only used their device for a few hours).

App-usage data

App usage logs were gathered between August 19th and September 23rd, 2013. The loggers ran for an average of 18 days on each participant's mobile phone. We logged a total of 32,524 app launches (M=2,323 per participant) Our 14 participants used 258 unique mobile apps over the course of the study.

RESULTS

In this section we present the results of our analysis of mobile device use, individual application use as well as the duration of those interactions across our three data sets.

Mobile Device Use

Past mobile usage studies have explored overall mobile device use in order to improve various interaction elements, such as application shortcuts, widgets and notifications on those devices. For example, ProactiveTasks [1] focused on understanding mobile device use durations in order to streamline short mobile device interactions. The authors propose a classification of mobile device use based on duration and interaction type and distinguish three types of device interactions: 1) *Glances*, in which the user simply turns on the screen of their device, checks information on their lock screen or home screen and then turns the screen off without interacting with any specific mobile applications; 2) *Reviews*, in which the user interacts with the device for at most 60 seconds and launches / interacts with at least one application during that time; and 3) *Engages*, in which the user interacts with the device for longer than 60 seconds during which the user also launches / interacts with at least one mobile application.

In this section we explore these three different types of device interactions within our three datasets. To conduct this analysis, we divide our datasets into a series of device sessions following the same method as in [1]: device use sessions start when the screen is turned on and end when the screen is turned off. When two sessions are 5 seconds or less apart, we merge them into the same session. Figure 1 shows the proportions of different device use sessions from the original ProactiveTasks formative study [1] along with our three different data sets. Note that the percentages in the figure are medians across users.

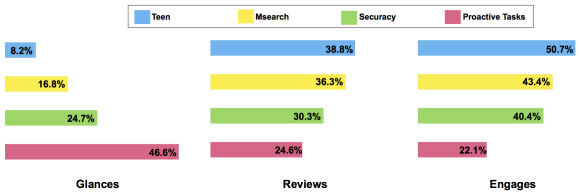

Figure 1. Comparison of Glance, Review and Engage interactions types across our three mobile data sets along with the formative ProactiveTasks study. The percentages presented represent the median across participants.

Overall we find a large difference between the distributions of session types across the four studies. The original study from Banovic et al. [1] reports the largest percentage of glance sessions followed by Securacy, MSearch and Teen studies in decreasing order. The Teen study has by far the least percentage of glance sessions (8.2%), that is where users turn on the screen of their mobile phone and check their lock screen or home screen but do not interact with any actual applications. The order is reversed when it comes to engage sessions with the Teen dataset having the largest proportion of engages (50.7%), followed by MSearch, Securacy, and the original ProactiveTasks study. Thus users in the Teen study interact with their device for longer periods ($>$ 60 seconds) and launch at least one application during those periods.

Application Use

Table 2 lists the top-10 applications across our three studies in terms of frequency and percentage of application launches. To conduct this analysis we count the total number of app launches on a per app basis for each dataset and then extract the top 10 of these apps in terms of volume of app launches[4].

We find that there are only 2 applications that appear in the top-10 list of all three datasets, namely: Facebook and Contacts. While the exact communications apps differ across the datasets, communications applications were prevalent across all three datasets. For example, messaging apps (e.g., SMS and WhatsApp) and voice-based communication applications (e.g., Skype, Viber and the built-in dialer) appear in all three top-10 lists. Music applications (e.g., Spotify and Pandora) appear in top-10 lists of both the Securacy and Teen datasets. Securacy and MSearch users appear to use browsers like Chrome more frequently that users in the Teen study.

Similar results were found in recent work by Ferreira et al. [17] which show Facebook, Contacts, and communications apps like SMS, Skype and WhatsApp in their top-10 app lists. Interestingly, despite the teens using games in their day to day life, no games apps appeared in their top 10 app list. This is mainly due to the fact that communications based applications were more popular among the teenagers in our study.

Understanding Micro-Usage

Mobile application micro-usage, a term coined by Ferreira et al. [17] is defined as brief bursts of interaction with applications. In their work, Ferreira et al. [17] found a natural break point in their data at approximately 15-seconds and thus defined micro-usage within their dataset as interactions lasting 15 seconds or less. They found that 41.5% of all application sessions were micro-usage. Note that an application session starts when the user launches an application and ends when the user exits the application or the screen is turned off. Thus we explored all application sessions across our 3 datasets to determine: 1) what the natural break point within those datasets are; and 2) what proportion of applications sessions account for micro-usage.

We followed a similar approach to Ferreira et al. [17] and used k-means clustering to determine natural break points in our three datasets. K-means clustering partitions our application sessions into k clusters in which each observation belongs to the cluster with the nearest mean. For the Securacy

[4]Note that the *MSearch App* listed in the top-10 list for the MSearch study is the study probe used in the MSearch study for collecting mobile search and browsing information.

	Securacy			MSearch			Teen		
No	**App**	**# App Launches**	**% Perc**	**App**	**# App Launches**	**% Perc**	**App**	**# App Launches**	**% Perc**
1	WhatsApp	4764	3.13	SMS	6217	10.74	SMS	6901	21.22
2	Facebook	2829	1.86	Facebook	6136	10.60	Phone	3324	10.22
3	Viber	1401	0.92	Phone	5157	8.91	Instagram	3150	9.69
4	Contacts+	1088	0.71	Built-in browser	3495	6.04	Snapchat	1649	5.07
5	Chrome	1071	0.70	Contacts	2980	5.15	Twitter	1222	3.76
6	Messenger	776	0.51	Email	2279	3.94	Facebook	1211	3.72
7	Firefox	716	0.47	Chrome	2040	3.53	Contacts	1133	3.48
8	Skype	706	0.46	Msearch App	1700	2.94	Email	978	3.01
9	Spotify	677	0.44	Gmail	1537	2.66	Pandora	786	2.42
10	Clash Of Clans	633	0.42	Instagram	1464	2.53	Calendar	774	2.38
	Total	**14661**	**9.62**	**Total**	**33005**	**57.04**	**Total**	**21128**	**64.96**

Table 2. Top 10 apps for each of our 3 datasets. # App Launches is the number or frequency of app launches for a given app in a given study, % perc is the percentage or proportion of app launches for a given app in a given study. The total shows the total number of app launches that these top 10 apps account for in the study in question.

dataset we found a natural break point at 16.6 seconds and found that 53.2% of all application usages were <=15 seconds. To compare, we found the natural break point to be larger in both MSearch and Teen datasets: 22.5 seconds and 21.5 seconds, respectively. If we look at the proportion of micro-usage within each of the datasets we find that approximately 48% of all application usages within the teen study were <=15 seconds, with almost 55% being 21.5 seconds or less. We find similar results for MSearch with > 48% of all application usages lasting 15 seconds or less, and approximately 56% lasting 22.5 seconds or less.

Figure's 2, 3 and 4 shows a probability distribution function of application usage in Securacy, MSearch and Teen studies. Specifically they show the probability of application sessions duration for the top-10 applications across all participants for each dataset. As previously reported by Ferreira et al., different applications exhibit different micro-usage behavior, and we found this across all three datasets.

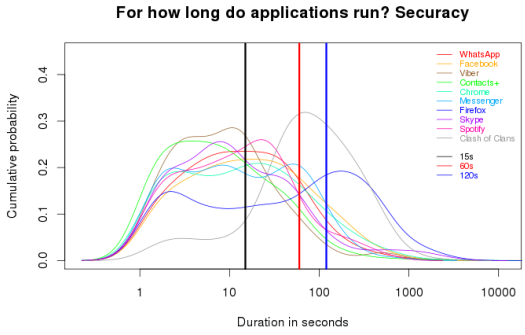

Figure 2. Probability of application session length in the Securacy study, across all users, regarding the top-10 applications

DISCUSSION & FINDINGS

Our analysis of mobile device usage patterns amongst our 3 diverse mobile usage datasets revealed individual differences in app usage behaviors in terms of individual application use, as well as the duration of those interactions. We found that our results differed not only from each other, but also from the prior formative studies we sought to revisit.

In studying the proportion of glance, review and engage sessions we found that the formative ProactiveTasks work had a much higher volume of glance sessions than any of our 3

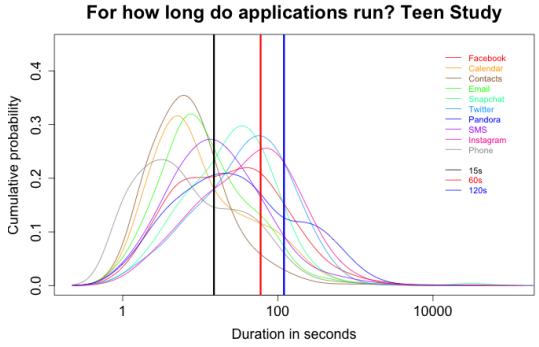

Figure 3. Probability of application session length in the Teen study, across all users, regarding the top-10 applications

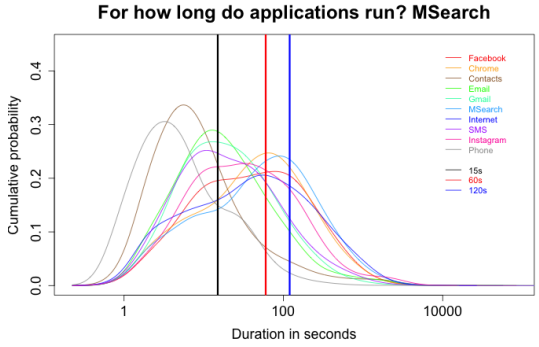

Figure 4. Probability of application session length in the MSearch study, across all users, regarding the top-10 applications

datasets (46.6% for ProactiveTasks vs. 24.7% for Securacy, 16.8% for MSearch and just 8.2% for the Teen study). These large differences in the usage patterns found could be due to a number of factors. All four data sets contained the necessary information to reconstruct the device use sessions and replicate both the ProactiveTasks and Micro-Usage analyses. Likewise all the four studies were conducted within one year of each other.

However, the ProactiveTasks formative study had by far the smallest number of participants (10), which makes it difficult to generalize findings from that study to the other three data sets. Also, the user populations across these four data sets are quite different. The ProactiveTasks formative study

recruited primarily college age mobile power users, the Securacy data set focused on security minded users of the Google Play Store, the MSearch study included adult participants who were active mobile searchers, and the Teen data set recruited participants from the teen population in the Bay Area. It is likely that each of these sub-populations have subtle differences in their goals when using their mobile devices. Although MSearch was probably the most representative of the general user population, much like the other three data sets, it still suffers from possible self-selection bias. It is also important to note that unlike the other data sets, the Teen data set contains both iPhone and Android users. Thus, even the subtle differences between the existing interaction elements across these two mobile platforms (e.g., iOS vs Android) could have caused differences in how much information the users are able to glean from their lock screen or home screen to make glances worthwhile.

Exploring the top 10 application lists across the datasets also revealed differences in behaviors. Interestingly the MSearch and Teen studies have the greatest overlap, with 6 applications appearing in both top-10 lists: SMS, Facebook, Phone, Email, Contacts and Instagram. Similarities between these two different data sets are perhaps due to both of these studies having participants in the Greater Bay Area and both using a similar mixed-method approach (i.e., combining device use logging along with qualitative insights from interviews and daily diaries).

We also find that the top-10 applications represent a large proportion of total application usage in both the MSearch and Teen datasets (57% and 65% respectively), however, they only account for 9.6% of all application usage in the Securacy dataset. This implies that there is a longer tail of application usage among the Securacy users, potentially attributed to the large geographic spread of users and application use in the Securacy study.

In replicating the Micro-usage analyses across our 3 new datasets we also found some differences in usage patterns. For example the natural break point found in the Securacy dataset of 16.6 seconds was more inline with the formative paper by Ferreira et al. [17] which reported micro-usage at 15 seconds. In contrast MSearch and Teen studies revealed a higher natural break point of 22.5 and 21.5 seconds each. In looking at the micro-usage patterns across the top-10 apps in each dataset we were able to validate that apps most likely to be micro-used were *focused-applications* (i.e., applications with limited functionalities or that are task-oriented), and *social-applications*: e.g., in Securacy's dataset: Viber, Skype, Contacts+, WhatsApp, Facebook. Least likely to be micro-used were *information seeking* applications and *leisure* applications: e.g., in Securacy's dataset: Spotify, Clash of Clans, Browsers (Chrome, Firefox). We found similar patterns in both the Teen and MSearch studies in terms of focused applications again being the most likely micro-used application, e.g., in the Teen study we see Calendar, Contacts, Email, Phone and SMS are micro-used, while in the MSearch study Phone, Contacts, Email, Gmail and SMS are micro-used. In the MSearch study information seeking applications like Chrome, Internet and our MSearch probe were the least likely micro-used applications.

While similar patterns of micro-usage were found across all three datasets in terms of focused applications, in the Teen and MSearch studies, we found that social-applications such as SnapChat, Twitter and Instagram (with the exception of Facebook) were not micro-used, a finding that is likely indicative of a difference in how the recruited participants of these two studies use these particular social applications.

Overall what is interesting is that some of the differences found were in fact consistent across the three datasets we analyzed. For example, the higher percentage of engage sessions in the Teen data set than in the Securacy data set appear to follow the trends we found regarding their respective top-10 apps, e.g., Facebook having smaller portion of micro-uses in the Teen data set than in the Securacy data set. Ultimately these findings point to potential differences in behavior between participants in the different studies rather than differences in how the data was collected or analyzed.

REFLECTIONS

In this section we reflect on deeper issues related to replicability and generalizability and highlight that while replicability and generalizability may not be feasible when studying such mobile device usage patterns, identifying and understanding the nuances of mobile usage among different user sub-populations with very different user behaviors and possibly even different user needs is important and does indeed help us as a community advance our field and our overarching understanding of what makes the mobile space so unique.

Replicability != Reproducibility

Replicability in research can be defined as the extent to which we can re-study a phenomenon and repeat the findings of an initial study. Part of this re-study is only possible if we have sufficient details about the procedures and methodologies employed in the first study. It's generally assumed that if we can achieve replicability, there is some level of consistency amongst our studies and we can be more confident in the validity of our findings.

As a community we do a good job at ensuring studies are *reproducible*. As can been seen in Table 1, most scientific papers include sufficient details regarding the participants we recruit and (sometimes) the sampling method employed; the duration, dates and deployment type of our studies; the logging and data collection mechanisms employed; the incentives provided (if any) and most importantly the analysis conducted. This helps ensure that at the very least, the data sets collected and the analyses conducted in most of the published studies can be reproduced. However, just because a study is reproducible does not mean that it's replicable [13]. In theory, a study is only replicable if you perform the exact same experiment or study (at least) twice; collect data in the same way both times, perform the exact same data analysis, and perhaps even arrive at the same conclusions.

In this paper we sought to compare three different mobile app usage datasets and to replicate past work on micro device us-

age to see if we could gain a more in-depth understanding of short, bursty mobile device interactions. However, what we found is that replicability simply isn't possible in these kind of mobile studies. We have shown through our case study that none of the datasets produced the same results as each other, nor as the original work we were attempting to compare them against, despite the fact that the three studies collected the same mobile usage data. Both the MSearch and Teen studies actually used the exact same application loggers and employed the same mixed-method approach, yet they still yielded differences in usage patterns. And while these differences might reflect new practices in micro-usage behavior, the issue is that we cannot attribute those differences to one specific factor or a set of factors.

We are a community of researchers who are scattered around the globe. We have access to different mobile devices and technologies, to different sets of people and different user populations; and to different resources both in terms of recruitment and incentives. The fast-pace of technological change makes every attempt at collecting mobile usage data a "one-shot" operation. When we set out to study some phenomena in the mobile space, we often do so with our own set of objectives and our own research questions in mind. Our study designs often reflect those objectives and are targeted towards answering those specific research questions. In addition the phenomena that we are trying to study is often not well understood until after we explore and analyze the data we collect, making it more difficult to design studies with replicability in mind.

We have shown that by combining these multiple datasets we have been able to look at mobile usage behaviors from multiple viewpoints. And there is clear value in doing this type of work because we can begin to build up a more complete, holistic picture of the diversity of mobile device usage. However, combining datasets will not help with replicability. Fundamentally, work of this nature is not replicable beyond the given population and instead should be interpreted as such.

Finally, it's important to note that reproducing such studies is important in our dynamic field. Just because a set of researchers in some part of the globe have already published a study understanding a set of mobile behaviors, doesn't mean that the problem or characterization in question is fully solved. Reproducing a study in a different part of the world, with a different user population and at a different point in time should be encouraged. Given the rapid pace of change in mobile, it actually represents one of the domains in which reproducing, and not necessarily replicating, studies should be encouraged, in particular if such work acknowledges its caveats and provides some additional interesting understanding of mobile behaviors.

Generalizability

Our case study has also led us to reflect upon the generalizability of our results. As researchers we have all likely received a review of our research approach, either commending or critiquing the generalizability of our study. As reviewers for the Mobile HCI community we may have even commented on the generalizability of fellow researchers work. Generalizability refers to the ability to apply results from a study sample to the larger population from which the sample was selected. While generalizability generally requires a sample population of interest, it also requires us to understand the contexts in which the study is conducted and how those contexts might impact on the results.

In our case we were trying to understand micro-usage behaviors across three different populations: Teens in the Bay Area, Adult Android users in the Bay Area who engage in mobile search, and security conscience Android users who use Google Play. While each study in isolation was well conducted, using similar on-device app logging software, with strong study design choices and interesting insights, the differences we found in the results show that generalizing across studies is challenging. For example, would the results of the teen study hold for teens in other areas in the US? Would the results from the MSearch study hold for iPhone searchers in the Bay Area? Would the results of the Securacy study hold for other Google Play users who are interested in lifestyle and entertainment related apps rather that security apps? The answer is probably not.

Fundamentally, this type of work is investigating socio-technical systems through a quantitative lens of investigating device and app usage patterns (frequency, durations, types of interactions). There is clear value in these types of work in uncovering interesting user practices and gaining insights about future areas of exploration. However, we as a community need to be careful about interpreting the results. And likewise as researchers we need to be careful about scoping our findings and resulting implications.

Unlike some areas of HCI where we are investigating rather fundamental properties of humans (such as the biomechanical properties represented with a Fitts' Law study), we are investigating snapshots of behavior with our participants and their mobile devices. As such, our findings are inherently tied to the participants of our studies. While we might present several statistics and quantitative findings about usage that are informative, it is unlikely that those statistics or findings would hold to another population. Similar to our issues with replicability, fundamentally this work is not generalizable beyond the given population and again should be interpreted as such.

The key difference between these types of mobile data usage studies and more qualitative HCI studies is that instead of gaining deep insights about the population through more subjective methods, these app logging studies are using quantitative techniques to gain different insights about that population. Even larger scale quantitative studies such as those conducted by Böhmer et al. [3] really should to be interpreted as specific to the participants selected.

Ultimately there is still great value in studying these individual user populations at specific points in time. This allows individual researchers and the community at large to combine and contrast the experiences collected from different contexts and user populations in order to build a better understanding

of different types of mobile users and the various behaviors that tie them together.

It is also important to note that replicability and generalizability should not always be the only goals when designing mobile device use studies. The differences in the device use studies show the importance of identifying different user sub-populations with very different user behaviors and possibly even different user needs. Lumping different kinds of users together in a user study with a large sample size might help in replicability and generalizability of findings, but it would hide the subtle differences between the sub-populations. This would in turn impact the value of any novel interventions designers create to address the needs of this broader population because although it would satisfy most users, it might not satisfy the specific needs of users within those sub-populations.

CONCLUSIONS

In this paper we used data collected from three diverse smartphone usage studies to replicate prior work on mobile application usage to see if we could add to our understanding of mobile device behaviors. Our studies were: a 2-week mixed-method study with 14 teens in the Bay Area; a 2-week mixed method study of 20 Adult users of mobile search in the Bay Area; and a larger-scale app market deployment with 87 security minded Google Play users. Our analysis pointed to a number of differences in the usage patterns across our three datasets but also in relation to prior related work. Specifically we found differences in the top-10 apps in each dataset, in the durations and types of interactions as well as in micro-usage patterns. Given that we could not attribute those differences to a specific factor or set of factors, we presented a deeper discussion and reflection on the challenges and caveats of conducting research of this nature. The key message for our community is that data of this nature provides great value in understanding how sub-populations of people use their mobile devices and can indeed be used to inform the design of novel interactions that improve mobile device interactions. However these insights come with boundaries that we much acknowledge when conducting and more importantly presenting work of this nature, in particular related to generalizability and replicability. We hope that our somewhat troubled analysis sparks discussion and debate within our community about how we conduct research of this nature and how we can continue to improve and evolve our methods and approach.

ACKNOWLEDGEMENTS

We would like to thank the Mobile HCI reviewers for their very useful comments which were used to improve this paper. We would like to thank all the participants who took part in our three studies of mobile application usage. We would also like to thank our colleagues involved in the Teen study (Frank Bentley, Beverly Harrison and Matt Rafalow) as well as colleagues involved in the MSearch study (Juan Pablo Carrascal). Denzil Ferreira is is partially funded by the Academy of Finland (Grants 276786, 285062, 286386), and the European Commission (Grants PCIG11-GA-2012-322138 and 645706-GRAGE).

REFERENCES

1. Banovic, N., Brant, C., Mankoff, J., and Dey, A. Proactivetasks: The short of mobile device use sessions. In *Proceedings of MobileHCI '14*, ACM (2014), 243–252.
2. Bentley, F. R., and Metcalf, C. J. Location and activity sharing in everyday mobile communication. In *CHI '08 Extended Abstracts*, ACM (2008), 2453–2462.
3. Böhmer, M., Hecht, B., Schöning, J., Krüger, A., and Bauer, G. Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In *Proceedings of Mobile HCI'11*, ACM (2011), 47–56.
4. Brandt, J., Weiss, N., and Klemmer, S. R. Txt 4 l8r: Lowering the burden for diary studies under mobile conditions. In *CHI '07 Extended Abstracts*, ACM (2007), 2303–2308.
5. Brown, B., McGregor, M., and McMillan, D. 100 days of iphone use: Understanding the details of mobile device use. In *Proceedings of the MobileHCI '14*, ACM (2014), 223–232.
6. Carrascal, J. P., and Church, K. An in-situ study of mobile app & mobile search interactions. In *Proceedings of CHI '15*, ACM (2015), 2739–2748.
7. Carter, S., Mankoff, J., and Heer, J. Momento: support for situated ubicomp experimentation. In *Proceedings of CHI '07*, ACM (2007), 125–134.
8. Cherubini, M., and Oliver, N. A refined experience sampling method to capture mobile user experience. In *International Workshop on Mobile User Experience Research held at CHI '09* (2009).
9. Church, K., Cherubini, M., and Oliver, N. A large-scale study of daily information needs captured in situ. *In Transactions on Human-Computer Interaction (TOCHI) 21*, 2 (2014), 10.
10. Consolvo, S., and Walker, M. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing 2*, 2 (4 2003), 24–31.
11. Dearman, D., and Truong, K. Evaluating the implicit acquisition of second language vocabulary using a live wallpaper. In *Proceedings of CHI '12*, ACM (2012), 1391–1400.
12. Do, T. M. T., Blom, J., and Gatica-Perez, D. Smartphone usage in the wild: A large-scale analysis of applications and context. In *Proceedings of the ICMI'11*, ACM (2011), 353–360.
13. Drummond, C. Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at ICML* (2009).
14. Falaki, H., Mahajan, R., Kandula, S., Lymberopoulos, D., Govindan, R., and Estrin, D. Diversity in smartphone usage. In *Proceedings of MobiSys '10*, ACM (2010), 179–194.

15. Ferreira, D., Dey, A. K., and Kostakos, V. Understanding human-smartphone concerns: a study of battery life. In *Pervasive Computing*, Springer-Verlag (Berlin, Heidelberg, 2011), 19–33.

16. Ferreira, D., Ferreira, E., Goncalves, J., Kostakos, V., and Dey, A. K. Revisiting human-battery interaction with an interactive battery interface. In *Proceedings of Ubicomp '13*, ACM (2013), 563–572.

17. Ferreira, D., Goncalves, J., Kostakos, V., Barkhuus, L., and Dey, A. K. Contextual experience sampling of mobile application micro-usage. In *Proceedings of MobileHCI '14*, ACM (2014), 91–100.

18. Ferreira, D., Kostakos, V., Beresford, A. R., Janne, L., and Dey, A. K. Securacy: An empirical investigation of android applications? network usage, privacy and security. In *Proceedings of the 8th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)* (2015).

19. Ferreira, D., Kostakos, V., and Dey, A. K. Lessons learned from large-scale user studies: Using android market as a source of data. *International Journal of Mobile Human Computer Interaction 4*, 3 (1 2012), 28–43.

20. Ferreira, D., Kostakos, V., and Dey, A. K. Aware: mobile context instrumentation framework. *Frontiers in ICT 2*, 6 (2015).

21. Fischer, J. E. Experience-sampling tools: a critical review. *Mobile Living Labs 9* (2009), 1–3.

22. Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., and Landay, J. A. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*, ACM (2007), 57–70.

23. Henze, N., Pielot, M., Poppinga, B., Schinke, T., and Boll, S. My app is an experiment: Experience from user studies. *Developments in Technologies for Human-Centric Mobile Computing and Applications* (2012), 294.

24. Intille, S. S., Rondoni, J., Kukla, C., Ancona, I., and Bao, L. A context-aware experience sampling tool. In *CHI '03 Extended Abstracts*, ACM (2003), 972–973.

25. Kamisaka, D., Muramatsu, S., Yokoyama, H., and Iwamoto, T. Operation prediction for context-aware user interfaces of mobile phones. *SAINT'09. Ninth Annual International Symposium on Applications and the Internet* (2009), 16–22.

26. Lee, U., Lee, J., Ko, M., Lee, C., Kim, Y., Yang, S., Yatani, K., Gweon, G., Chung, K.-M. . M., and Song, J. Hooked on smartphones: An exploratory study on smartphone overuse among college students. In *Proceedings of CHI '14*, ACM (2014), 2327–2336.

27. McMillan, D., Morrison, A., Brown, O., Hall, M., and Chalmers, M. *Further into the Wild: Running Worldwide Trials of Mobile Systems*. Springer Berlin Heidelberg, 2010, 210–227.

28. Oulasvirta, A., Rattenbury, T., Ma, L., and Raita, E. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing 16*, 1 (1 2012), 105–114.

29. Palen, L., and Salzman, M. Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of CSCW '02*, ACM (2002), 87–95.

30. Pejovic, V., and Musolesi, M. Interruptme: Designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the UbiComp '14*, ACM (2014), 897–908.

31. Pielot, M. Large-scale evaluation of call-availability prediction. In *Proceedings of UbiComp '14*, ACM (2014), 933–937.

32. Pielot, M., Church, K., and de Oliveira, R. An in-situ study of mobile phone notifications. In *Proceedings of MobileHCI '14*, ACM (2014), 233–242.

33. Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., and Aucinas, A. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of UbiComp '10*, ACM (2010), 281–290.

34. Raento, M., Oulasvirta, A., Petit, R., and Toivonen, H. Contextphone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing 4*, 2 (4 2005), 51–59.

35. Rahmati, A., and Zhong, L. Studying smartphone usage: Lessons from a four-month field study.

36. Ramanathan, N., Alquaddoomi, F., Falaki, H., George, D., Hsieh, C., Jenkins, J., Ketcham, C., Longstaff, B., Ooms, J., Selsky, J., Tangmunarunkit, H., and Estrin, D. ohmage: An open mobile system for activity and experience sampling. In *PervasiveHealth*, IEEE (2012), 203–204.

37. Shepard, C., Rahmati, A., Tossell, C., Zhong, L., and Kortum, P. Livelab: Measuring wireless networks and smartphone users in the field. *ACM SIGMETRICS Performance Evaluation Review 38*, 3 (2011).

38. Shin, C., Hong, J.-H. . H., and Dey, A. K. Understanding and prediction of mobile application usage for smart phones. In *Proceedings of UbiComp '12*, ACM (2012), 173–182.

39. Srinivasan, V., Moghaddam, S., Mukherji, A., Rachuri, K. K., Xu, C., and Tapia, E. M. Mobileminer: Mining your frequent patterns on your phone. In *Proceedings of UbiComp '14*, ACM (2014), 389–400.

40. Truong, K. N., Shihipar, T., and Wigdor, D. J. Slide to x: unlocking the potential of smartphone unlocking. In *Proceedings of CHI '14*, ACM (2014), 3635–3644.

41. Vaish, R., Wyngarden, K., Chen, J., Cheung, B., and Bernstein, M. S. Twitch crowdsourcing: crowd contributions in short bursts of time. In *Proceedings of CHI '14*, ACM (2014), 3645–3654.

42. Vastenburg, M. H., and Herrera, N. R. Adaptive experience sampling: addressing the dynamic nature of in-situ user studies. In *Ambient Intelligence and Future Trends-International Symposium on Ambient Intelligence (ISAmI 2010)*, Springer (2010), 197–200.

43. Vetek, A., Flanagan, J. A., Colley, A., and Keränen, T. Smartactions: Context-aware mobile phone shortcuts. In *Proceedings of INTERACT '09*, Springer-Verlag (2009), 796–799.

44. Wagner, D. T., Rice, A., and Beresford, A. R. Device analyzer: Large-scale mobile data collection. *SIGMETRICS Performance Evaluation Review 41*, 4 (4 2014), 53–56.

45. Yan, T., Chu, D., Ganesan, D., Kansal, A., and Liu, J. Fast app launching for mobile devices using predictive user context. In *Proceedings of MobiSys '12* (6 2012), 113–126.

46. Zhang, C., Ding, X., Chen, G., Huang, K., Ma, X., and Yan, B. *Nihao: A Predictive Smartphone Application Launcher*. Springer Berlin Heidelberg, 2013, 294–313.