A Systematic Assessment of Smartphone Usage Gaps

Niels van Berkel, Chu Luo, Theodoros Anagnostopoulos, Denzil Ferreira,
Jorge Goncalves, Simo Hosio, Vassilis Kostakos
Center for Ubiquitous Computing
University of Oulu, Finland
{niels.van.berkel, chu.luo, tanagnos, denzil.ferreira,
jorge.goncalves, simo.hosio, vassilis}@ee.oulu.fi

ABSTRACT
Researchers who analyse smartphone usage logs often make the assumption that users who lock and unlock their phone for brief periods of time (*e.g.*, less than a minute) are continuing the same "session" of interaction. However, this assumption is not empirically validated, and in fact different studies apply different arbitrary thresholds in their analysis. To validate this assumption, we conducted a field study where we collected user-labelled activity data through ESM and sensor logging. Our results indicate that for the majority of instances where users return to their smartphone, *i.e.*, unlock their device, they in fact begin a new session as opposed to continuing a previous one. Our findings suggest that the commonly used approach of ignoring brief standby periods is not reliable, but optimisation is possible. We therefore propose various metrics related to usage sessions and evaluate various machine learning approaches to classify gaps in usage.

Author Keywords
Mobile devices; session; classification models; human behaviour; phone usage; ESM; machine learning.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION
Recent studies on daily user interaction with smartphones have led to an increased understanding of how users use these popular devices, and how manufacturers and designers can further improve these devices. An important element of a user's interaction with their phone is the completion of a wide variety of purpose-driven objectives (*e.g.*, call someone, complete an achievement in a mobile game, check e-mail). These objectives can range from brief tasks confined within a certain application to overarching tasks spanning multiple applications and services. Additionally, it is possible to group objectives into usage sessions, and one session can contain multiple objectives.

The richness of functionality and interaction that smartphones offer has been increasingly used as a proxy to study and quantify human behaviour [4,12,27,33]. For example, analysing which applications a person uses may be indicative of lifestyle choices. In literature, such an analysis considers *application sessions,* typically defined as a continuous period of time in which an application is both active and visible [5,12,31]. It can also be insightful to study people's overall use of their phone, regardless of specific applications. Surprisingly, literature does not provide a clear definition for *phone usage sessions,* and in fact many definitions exist that are often ambiguous or based on assumptions [4,5,29]. For example, Carrascal & Church [5] define a usage session as a sequence of actions during which the display was not turned off for more than 30 seconds. This 30-second threshold is arbitrary, and as a result the authors note that user goals often spanned multiple sessions [5]. So far, no study has empirically investigated and quantified *phone usage sessions*, and considered whether researchers should ignore brief timeouts or signal the beginning of a new usage session.

In this paper, we conduct an empirical investigation of phone usage sessions that combines automated data logging and user-provided labelling. We use the Experience Sampling Method (ESM) [21] to collect users' labels at the start of a phone usage session (*i.e.*, as the user unlocks the phone). We also unobtrusively gather interaction data (*e.g.*, screen status, application launches) from our participants during a 1-week long field deployment. Besides the ESMs, we do not introduce other changes to participants' everyday use of their device. From the collected data, we are able to empirically identify gaps in phone usage and examine heuristics that can help in answering the question: *should researchers ignore a particular gap when considering usage sessions?* Our paper contributes to the available corpus on everyday smartphone usage, specifically the analysis of usage sessions, and can provide benefits for mobile phone users, *e.g.*, support intermittent application usage [3], better battery usage predictions [11], or provide visual cues for incomplete tasks [22].

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858348

RELATED WORK

Falaki *et al.*'s [10] work pioneered the analysis of how users use their smartphones in daily life, focusing on the number and duration of user interactions, application usage, and generated network traffic. The researchers identify "interaction intervals" as a valuable information source – specifically for increasing devices' battery life: mobile phone interactions are mainly brief, with a few longer exceptions throughout the day. Furthermore, the authors model several characteristics of smartphone usage for the whole user population – though the model parameters may differ between users. Relevant to our work is the timeout between interaction intervals, modelled according to the Weibull distribution. This model suggests that the shorter the timeout between the previous interaction and now, the higher the chance for the next interaction to occur.

In a large-scale study, Böhmer *et al.* [4] report an average device usage of 59.23 minutes per day, with an average of 71.56 seconds spent in an application. Yan *et al.* [35] found that the time between screen unlock and subsequent screen lock is less than 30 seconds in half of the total instances. Pielot *et al.* [26] report an average of 63.5 incoming mobile notifications per day, mainly messages and emails. These results show that mobile phones are frequently used throughout the day, with a focus on short bursts of interaction. Ferreira *et al.* [12] describe this characteristic of phone usage as "application micro-usage" to describe extremely brief application sessions, and as the "checking habit" [25] due to the repetitive inspection of dynamic content on the user's smartphone.

Others focus on categorising types of mobile phone usage: *glance*, *review*, and *engage* [3], where a glance denotes the situation where the user only looks at the homescreen or lockscreen of the device, a review represents a brief interaction (less than a minute) with one or more applications, and engage describes longer lasting interactions (*i.e.*, for longer than one minute) in which the user uses one or more applications. Hintze *et al.* [16] distinguish between locked and unlocked usage sessions, as locked usage sessions offer only a limited range of functionalities, but more easily within reach (*e.g.*, checking the time, battery status, taking picture).

Phone usage sessions

Analysis of usage sessions has also been an active research area in the field of information retrieval. Jansen *et al.* define a *usage session* as "*a series of interactions by the user toward addressing a single information need*" [18]. Jones & Klinker [20] propose a hierarchy of goals, missions, and sessions, whereby a mission can be composed of multiple goals and a session can in turn constitute multiple missions. In this definition, the authors describe a usage session as "*all user activity within a fixed time window*" [20], or simply a slice of the user's time. According to the cited work, a session is therefore either composed of one or more user objectives, or constitutes a (continuous) period of time.

Studies investigating phone usage sessions have largely adopted these definitions from the field of information retrieval – sometimes including modifications to account for phone-specific use cases. One example is the distinction between an *application* usage session and a *smartphone* usage session [31]. An application usage session is the time spent using an application in the foreground - whenever the user switches to a different application, a new application usage session commences. A smartphone usage session is the combination of one or more application usage sessions (depending on a threshold value of potential idle/standby time between these application usage sessions). Carrascal & Church [5] define a session as an interaction sequence without turning off the display for more than 30 seconds, thus following [31] but apply a 30-second threshold. Böhmer *et al.* [4] use an identical definition, but refer to this as an 'application chain.' It is worth pointing out that these definitions exclude interaction with the lock-screen (*e.g.*, checking the time or glancing at notifications) [16].

Also common is the definition of a phone usage session based on active screen usage, where we consider the time between screen on and screen off (either through user action, or automatic idle timeout) as one session (*e.g.*, [10,16,25]). Both definitions consider a user's task as a set of running multiple applications, *e.g.*, looking up the address of a point of interest on a website after which the user uses a navigation application to reach the location.

Voice calls are a special case [16]: an incoming call activates the device's screen, regardless of the owner's presence or the call status (*i.e.*, answered, unanswered); on outgoing calls, for the majority of phones, the screen turns off when the user raises the phone to their ear to prevent accidental interference. According to Hintze *et al.* [16], calls account for 12.7% of the screen state switching.

Lastly, external factors can also interrupt a phone usage session, for example looking up to avoid collision when crossing the street. Real-world interruptions can lead to the device being temporarily turned off or put away in the pocket, either by the user control or automatically after the device's timeout margin.

Towards consistent terminology and analyses

Our systematic literature review on the definitions of phone usage sessions and application sessions revealed several divergent definitions. The majority of these definitions do not always include common smartphone use-cases (*e.g.*, missed incoming call, outgoing call, phone reboots), or are study-specific (Table 1). Furthermore, these definitions do not take into account the user's tasks and goals. Instead, technical mechanics of interaction form the base of these definitions (*e.g.*, turning on the screen).

To overcome the inconsistency in terminology, we propose a coherent model and terminology for describing smartphone usage, taking also into account phones with a lock-screen enabled. We show the set of possible

smartphone states and transitions between those states in Figure 1, and in Figure 2 we show a visual summary of application sessions and usage sessions. Our model distinguishes between using the phone in locked and unlocked condition. A *locked phone usage session* consists of the user interacting with the "lock screen" of their phone. An *unlocked phone usage session* consists of the user unlocking their phone and interacting with it.

For the remainder of our work, we primarily focus on identifying phone usage sessions as defined in Figure 2. Previous researchers have assumed that *briefly* entering the locked-display-off or the power-off state should not signal the end of a usage session, effectively ignoring these state changes – as long as the user returns to an active usage state within a certain time threshold (henceforth, T). This assumption appears to be reasonable, since previous work noted that a usage session may be interrupted without the actual user's intent to end the current session [9,27,30]. Therefore, such brief interruptions should not account for a new session.

However, in previous work there is no consensus on what the threshold T should be. Soikkeli *et al.* [31] use both T=0 and T=30 seconds, which lead to significantly different usage statistics: 20 vs 13 sessions a day with an average length of 4:23 minutes and 7:09 minutes, respectively. This highlights the main shortcoming of current literature, since using different thresholds results in drastically different findings. Both Church *et al.* [6] and Banovic *et al.* [3] use T=5 seconds. Böhmer *et al.* [4] define a session (or *'chain of app usage'*) as "*a sequence of apps that are used without the device being in standby mode for longer than 30 seconds*." Carrascal & Church [5] state "*we define a session as a sequence of interactions that occur without the device being in standby mode, i.e. the display switching off, for longer than 30 seconds.*" These definitions vary substantially, but more critically their characteristics are not empirically derived: they are simply intuitive.

Scope	Definition	Limitation(s)	Ref.
Phone usage session (also called *interaction session*, and *app chain*)	"*Combination of one or more application usage sessions (depending on threshold value of idle time in between these application usage sessions). That is, a group of application sessions with time interval less than T.*"	Lacking non-application based usage (*e.g.*, glancing at notifications during the locking state).	[31]
	Active screen time (*e.g.*, "*an interaction is defined as the interval that an application is reported to be on the foreground*" [10]).	Various phone events may activate the screen without user intent of actual device usage (*e.g.*, active phone ringing, OS notification or alarms, charging events).	[10,16,24,25]
	An interaction sequence without the device going into standby mode for more than 30 seconds.	30 second delimiter not based on any actual evidence.	[4,5]
	A non-voice session is a series of consecutive screen-on time (two minutes or more). The authors designated phone calls as voice sessions.	A non-voice session is a series of consecutive screen-on time (two minutes or more). The authors designated phone calls as voice sessions.	[29]
	"*[…] the duration that the LCD backlight was enabled less the time that the user was not interacting with the device (and resetting the idle time).*"	-	[24]
	The set of applications that were used between unlocking and locking the phone.	Lacking non-application based usage (*e.g.*, glancing at notifications during the locking state).	[19]
Application session	The time spent using an application in the foreground.	-	[31]
	"*[…] when, for how long and which applications were active and visible to the user.*"	-	[12]
Micro app. usage session	Application usage that lasts up to 15 seconds.	-	[12]
Locked / unlocked usage session	Locked usage sessions occur when device interaction takes place while the device remains locked by a keyguard (such as PIN, password, pattern, face unlock, fingerprint, or swipe to unlock).	Various phone events may activate the screen without user intent of actual device usage (*e.g.*, active phone ringing, OS notification or alarms, charging events).	[16]

Table 1. Summary of the various definitions found in literature to describe smartphone and application usage

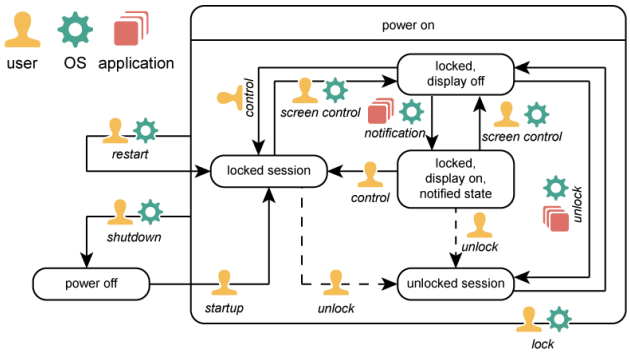

Figure 1. State transition diagram showing how actions by different actors may trigger changes to the smartphone state – dashed lines indicate possibility of a new user task

To the best of the authors' knowledge, no previous study has attempted to empirically validate different values of the threshold T when quantifying and categorizing differences in phone usage sessions. As a result, different researchers adopt different values, leading to largely incomparable results across studies. Here, we attempt to empirically derive a threshold T with the help of combined user labelling and sensor logging technique.

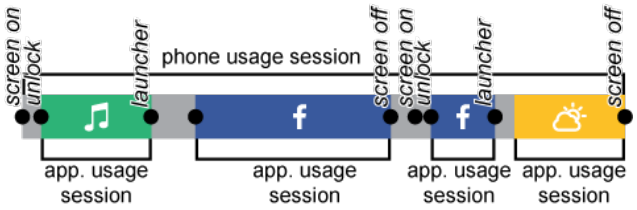

Figure 2. Visualisation of typical phone usage session, including four distinct application usage sessions

EXPERIMENTAL SETUP

We collected our data using a plugin developed for the AWARE framework [13], running continuously in the background of the participants' own Android phones. We collect the following:

- **ESM answer**: participants' answer to the ESM question (Figure 3): Why did you start using your phone? (Continue previous objective / Start on a new objective).
- **ESM status**: user's choice to reply, ignore or dismiss the ESM question.
- **Phone status**: various phone-related details (*e.g.*, phone state: reboot, shutdown; screen state: on, off, locked, unlocked; battery state: charging, discharging, current battery level).
- **Application names**: application launches and any notifications they trigger.

The plugin stored data upon a state change in the aforementioned data elements, and contained both a unique random ID per participant, and a timestamp. Furthermore, the plugin presents event-contingent ESM questions as popups as soon as participants unlock their screen. This allows us to collect data directly at the onset of phone usage, as opposed to allowing the participants to answer these questions at a later time (*e.g.*, defer them as a notification). Participants without a locking mechanism on their device receive the ESM question directly after turning on the device's screen (*i.e.*, entering the "unlocked session" state in Figure 1). Notifications are automatically dismissed when the user or OS either lock the phone or turn of the screen. Participants are able to dismiss the ESM message using the 'back' button on their phone.

The participants did not receive any other ESM questions during the study. Because of the technical nature and ambiguity of the term 'session,' we decided to avoid this phrase in the formulation of the ESM question. After considering a large number of alternatives, we decided to use the term 'objective,' given its definition of an action of short- to mid-term duration, with a focus on a specific action [18,20]. Also, we framed the ESM question to refer to the present unlocking action of the participant, rather than the most recent locking action (*e.g.*, "Why did you lock your phone the last time you locked it?"). We thus minimise the reliance on the participants' ability to recall from memory, reducing retrospection bias [8,23].

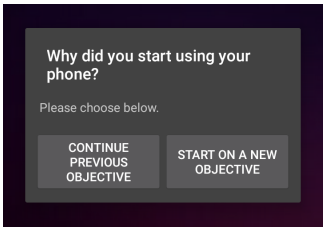

Figure 3. Question presented upon phone unlock

We collected data for seven days, as recommended by Hektner *et al.* [15] for the ESM method, to gather data from both weekdays and weekends, but also to avoid data degradation that occurs in longer studies. Given the high amount of notifications sent to the participants during the study – basically every time they use the phone – we decided to not extend duration beyond these seven days.

According to data collected from more than 17,300 BlackBerry users [24], half the number of mobile phone interactions take place within 90 seconds of each other. This rapid onset of successive interactions led us to decide not to set any inter-notification limit for issuing the ESM question. While an inter-notification limit would reduce participant strain by not asking them to answer the question on *every* single unlock, we would be unable to include those specific usage events in our analysis.

RECRUITMENT AND EXPERIMENTAL PROCEDURE

We recruited seventeen people from mailing lists of our university (13 males, 4 females; ages: 23-39 years old, M=26). The only requirement for participation was for participants to own an Android-based smartphone. Participants had a diverse range of educational backgrounds (*e.g.*, Economics, Computer Science, Linguistics, and Anthropology).

Procedure

To guarantee that participants understood what the ESM question asked of them, we held an individual training session for each participant and provided various examples of objectives (*e.g.*, text a family member with event details, find a nearby Italian restaurant and start route finding) to provide mental hooks for participants throughout the study duration. Furthermore, we discussed practical situations with participants in order to reach a common understanding of either continuing or starting on a new objective.

One example we provided to participants concerned the usage of a contacts application to add the name and contact details of a new acquaintance. This example covered several possible scenarios such as interruption by a third person, leading to an automatic phone lock (*continuing* objective if participant resumes task after interruption). A second practical example we discussed was the use of an instant messaging application while cooking dinner: participants resuming application usage after interruption (adding ingredients, stirring, etc.) should report this as *continuing* an objective.

We provided additional examples to further clarify what *new* and *continuous* objectives were when deemed necessary. Finally, to reduce ambiguous scenarios in which participants might believe to be 'multitasking' multiple objectives, we introduced the rule that only *directly* preceding objectives could be continued and that a user could only have one primary objective at a time. We offered participants the opportunity to ask any questions they might have about the described task.

Following this training session, we explained the functionality and data logging capabilities of the application, after which we installed the application on the participant's personal device. The deployment lasted for seven days and concluded with a one-on-one debriefing session. During the debriefing, we inquired about any potential problems the participants might have encountered, and requested participants to complete a short questionnaire. Finally, we removed the study software from the participants' phones, and participants received a compensation for their efforts (a cinema ticket).

ANALYSIS

We first coded the interaction data into usage sessions as defined in Figure 2. The analysis begins by considering uninterrupted usage sessions (*i.e.*, T=0), because for each such session we had an ESM label provided by participants. Participants' labels where used to characterise each usage session either as a *continuous* session or as a *new* session. The following features were associated with each session:

- **Application pattern**: the set of applications used within the session.
- **Categories pattern**: the set of categories of applications used in that particular session, as obtained from Google Play and categorized according to [5].
- **Day**: the day of the week in which the session occurred.
- **Hour**: the hour of the day in which the session occurred.
- **Gap**: the time (in milliseconds) between the end of the previous usage session and the beginning of the current usage session.
- **Label**: the label that the participant gave to this session via ESM. The value could be 0 (continuous) or 1 (new).

Session Classification Models

Previous work has adopted the use of a threshold T for deciding whether researchers should ignore a gap between two usage sessions and thus assume that the second session is a *continuous* session. This approach is conceptually identical to using a Constant Classifier, and thus follows the current common practice of using an arbitrary fixed threshold. We therefore built a Constant Classifier that takes as input a constant threshold T (in milliseconds) and classifies a usage session as a *continuous* session if the time attribute is less than T, or as a *new* session otherwise.

An alternative approach is to adopt a Similar Sets Classifier, which assesses the dynamics between two subsequent sessions. This approach assumes that there is a higher similarity between the attributes of two consecutive sessions if the second session is a *continuous* session. We measured similarity by means of a set similarity distance metric based on how many categories of applications the sessions share. We computed the distribution of the

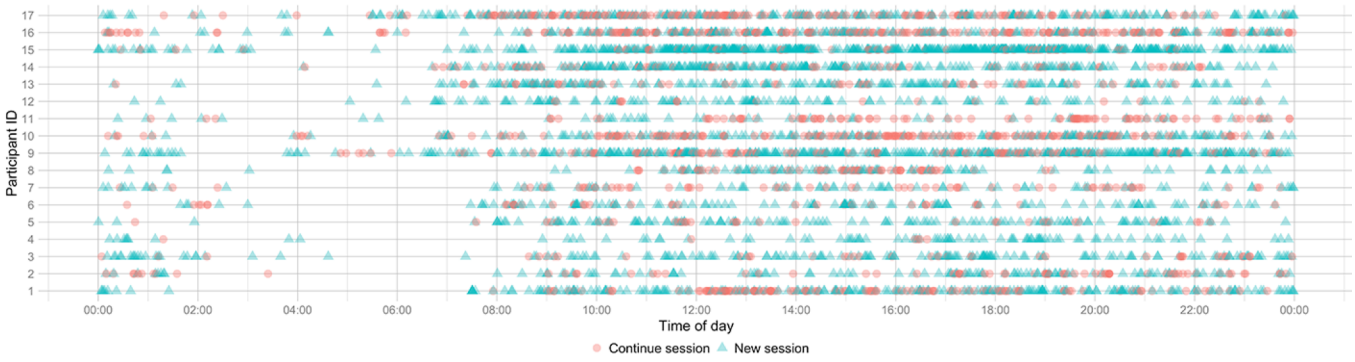

Figure 4. Overview of all ESM responses, plotted by time of day

distance metric for *new-continuous* session pairs and for *new-new* session pairs. If for an arbitrary pair of sessions their similarity distance metric is close to the *continuous* distance distribution, we classify the session pair as *new-continuous*, otherwise as *new-new*. This classifier requires information from both the preceding and following session, therefore being offline in nature.

Lastly, we also tested the use of the One Rule Classifier in WEKA [14], as proposed by [17]. This classifier resulted in the highest accuracy of all potential WEKA classifiers. The One Rule Classifier selects the minimum-error attribute and uses this attribute for classification. In the collected dataset, the minimum-error attribute is the hour attribute (*the hour of the day in which the session occurred*).

Results

During the study, the software triggered 5,397 ESM notifications, of which the participants answered 4,569 (average response rate of 83.78%, SD = 10.78), yielding a high response rate. The ESMs are also answered quickly (median 2 seconds, mean of 2.70 seconds and SD of 1.84 seconds after removal of outliers), suggesting a low burden to our participants.

Of all the ESM responses, 67.13% claimed to start a new objective and 32.90% to continue a previous objective. Due to Android's fragmentation and device-specific incompatibilities, seven participants had intermittent application name data. This has no impact in our data analysis however, since we successfully captured every time they lock and unlock their devices and their ESM answers regarding starting a new, or continuing a previous objective. We observe that participants start a new session much more often than a continuous session (Table 2). We found no significant correlation between the number of phone unlocks (i.e., total number of ESMs issued) and the ratio of the provided user answers ($r = -0.21$, $p = 0.41$), indicating that it is unlikely that our results are methodologically biased. Figure 4 shows an overview of the timing of participants' answer over the course of the entire week, plotted by time of day. As expected, we observe that most participants respond across working hours, and late night / early morning hours are less active. This plot also demonstrates the temporal granularity and breadth of the collected the ESM responses.

ID	Response rate	Answered ESMs	Continuous / New
1	95.18%	434	0.46 / 0.54
2	79.08%	155	0.26 / 0.74
3	67.31%	416	0.55 / 0.45
4	88.33%	227	0.24 / 0.76
5	94.23%	196	0.20 / 0.80
6	87.87%	507	0.27 / 0.73
7	96.86%	339	0.56 / 0.44
8	93.18%	328	0.39 / 0.61
9	71.13%	170	0.42 / 0.58
10	91.28%	136	0.68 / 0.32
11	84.24%	155	0.42 / 0.59
12	75.96%	139	0.45 / 0.55
13	85.88%	657	0.10 / 0.90
14	97.25%	318	0.23 / 0.77
15	78.54%	161	0.22 / 0.78
16	61.90%	91	0.07 / 0.93
17	76.09%	140	0.11 / 0.89
Avg.	**83.78%**	**269**	**0.33 / 0.67**

Table 1. Overview of participant responses to ESM

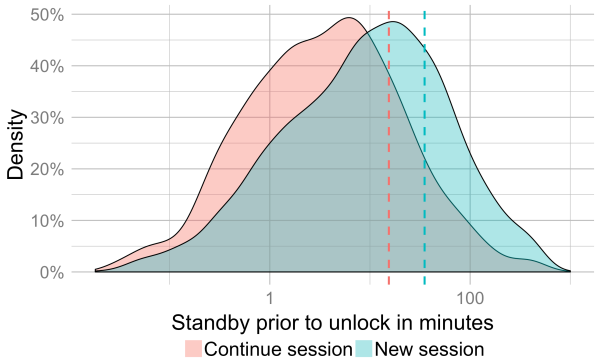

Figure 5. Probability of standby time prior to a usage session. Dashed lines indicate mean values

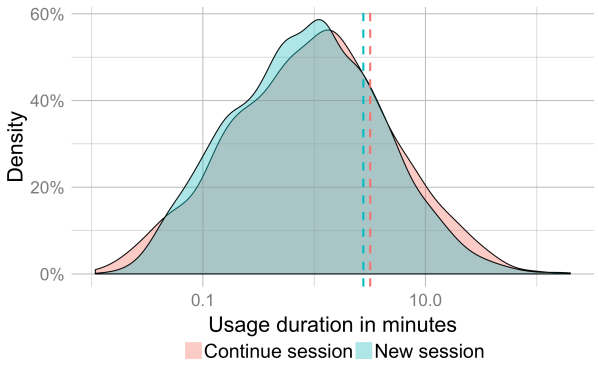

Figure 6. Probability of phone usage duration for interaction segments. Dashed lines indicate mean values

Figure 5 shows a density plot of the gap duration before a *continuous* (red) or *new* (blue) usage session. Mean gap duration prior to phone unlock is 15:26 minutes for *continuous* sessions, and 35:06 minutes for *new* sessions. We observe that both follow a Poisson distribution with varied skewness and kurtosis.

Figure 6 shows a density plot of the duration of sessions, which also follow a similar Poisson distribution. Recorded

mean usage duration is 2:46 minutes for *new* sessions and 3:11 minutes for *continuous* sessions. We mark (in dashed vertical lines) the mean combined duration of these consecutive sessions. Across all instances of use (which combines a *new* session with any number of subsequent *continuous* sessions) the average duration of use is 4:43 minutes. Note that participants sometimes labelled multiple consecutive usage sessions as break sessions. In Figure 7, we show how often participants labelled two or more consecutive sessions as *continuous* sessions.

As phone notifications potentially prompt users to unlock their phones, we further analyse participants' delay until they unlock their phone following a received notification. Table 3 shows the percentage of phone unlocks which occurred within different time frames following a received notification. We observe sharp differences between participants. For example, participant P5 generally responds quickly to notifications, *i.e.*, unlocking their phone shortly after receiving a notification. In contrast, the arrival of notifications did not affect the behaviour of other participants (*e.g.*, participants P6, P7). However, we did not find a significant effect between participants' tendency to respond quickly to an incoming notification and their ratio of *continuous* vs *new* sessions ($r = -0.32$, $p = 0.23$ for the 0-120 seconds bin). In other words, receiving more notifications leads to neither more *continuous* nor *new* sessions.

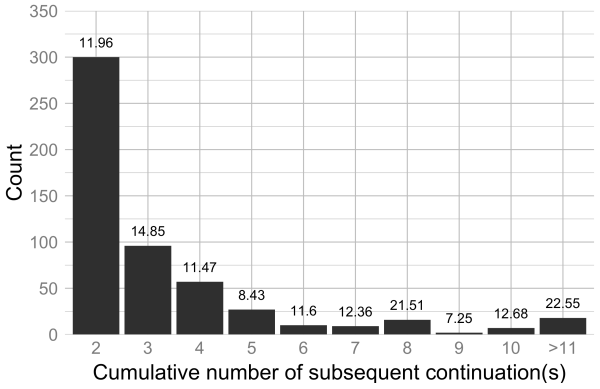

Figure 7. Histogram showing how often participants labelled two or more consecutive usage sessions as continuous sessions. Bar values indicate average duration (minutes) for each bin

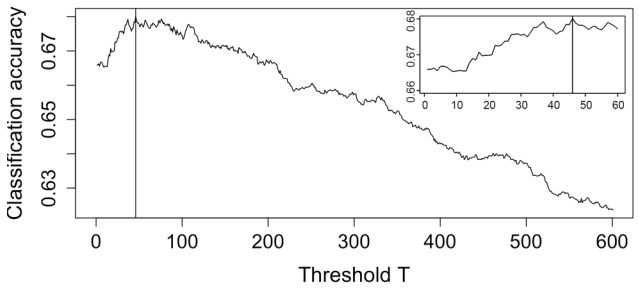

Figure 8. For each value of threshold T (x-axis, in seconds), we show the classification accuracy in determining whether a session is a continuous session or new session

Classifier Evaluation

Figure 8 shows the prediction accuracy for the Constant Classifier. We tested a range of values for threshold T, and we observe that for values below 30 seconds and above 120 seconds, the prediction accuracy degrades. The Constant Classifier performs best when the threshold is set to 45 seconds (accuracy=68%). For values of T at 5, 30, or 60 seconds, the classifier performs worse. The Similar Sets Classifier accuracy is 64.21%, while the One Rule classifier accuracy is 62.11%. We noted that the variability across participants is narrower in the Similar Sets classifier (min: 48.81%, max: 77.16%). The mean value of the similarity distance distribution for *continuous* sessions is 0.46076, and for *new* sessions is 0.76785. In some cases, the One Rule Classifier achieves better accuracy than the other classifiers (e.g. for participants P1, P2, P5, and P6), but on average the One Rule Classifier performed the worst.

We trained the three classifiers using user-labelled data – using a total of 10 out of 17 collected datasets to validate the classifiers using 10-fold cross validation. These datasets contain the collected data for each participant, seven datasets did not have sufficient data for validation. McNemar's chi-square statistic indicates a significant difference compared to the ground truth for all three classifiers with $p < 0.01$. Table 4 shows the performance values for the three classifiers.

	Time between receiving a notification and unlocking the phone				
ID	**0-15 s**	**0-30 s**	**0-60 s**	**0-90 s**	**0-120 s**
1	5.60%	9.80%	13.45%	16.81%	17.93%
2	3.38%	5.91%	8.44%	10.97%	11.81%
3	2.86%	4.29%	4.76%	4.76%	5.24%
4	6.96%	6.96%	6.96%	6.96%	6.96%
5	28.46%	29.64%	30.04%	30.56%	30.83%
6	0.72%	2.16%	2.16%	2.16%	2.16%
7	0.93%	0.93%	0.93%	0.93%	0.93%
8	1.21%	2.42%	2.42%	2.42%	2.42%
9	15.63%	18.30%	24.70%	27.98%	30.06%
10	2.28%	3.13%	3.13%	3.42%	3.42%
11	5.19%	5.84%	8.44%	8.44%	8.44%
12	14.40%	14.40%	16.00%	16.00%	16.80%
13	12.50%	12.50%	14.02%	14.77%	15.15%
14	11.96%	15.64%	21.47%	23.31%	26.38%
16	8.32%	9.81%	11.59%	12.48%	14.12%
17	4.40%	6.50%	7.97%	8.60%	8.81%

Table 2. For each participant we calculate the percent of received notifications that led to a phone unlock event within the specified time period (Participant 15 misses notification data)

Discussion

Previous work heavily relies on the *use of phone standby time analysis* to distinguish between *continuous* and *new* usage sessions (Table 1). We found that this technique alone is unable to make this distinction. We can only speculate that researchers have adopted this approach because intuitively, they assume that if a user is going to continue a task, then they are likely to do it after a brief gap in time. However, the assumption is a *fallacy of the converse*: our experiment shows that brief gaps in usage are very frequently a prelude to new sessions as well, and therefore a brief gap is not necessarily followed by a continuous session.

Classifier	Accuracy	Precision	Recall
Constant	68.0%	70.6%	95.6%
Similar Sets	64.2%	66.7%	93.1%
One Rule	62.1%	64.5%	90.1%

Table 3. Classifier performance

With our analysis of the usage sessions of participants, and by collecting participant labels, we are able to reliably establish a ground truth distinction between *continuous* and *new* usage sessions that is time independent. Specifically, Figure 5 shows that there is a considerable overlap in the duration of gaps preceding *continuous* and *new* phone usage sessions, making them effectively indistinguishable when considering time alone.

For instance, adopting a T=30 seconds threshold, as used in [4] and [5], actually only captures 30.37% of all true continuous sessions in our dataset. Additionally, our results show that 50.85% of gaps shorter than 30 seconds actually lead to a new session instead of a continuous session. This means that even when the phone is briefly on standby, users still frequently start a new objective after unlocking it. As one participant mentioned in the debriefing, "*Sometimes I just check my phone shortly and sometimes I use it for longer periods (Facebook etc.).*" (P12). Consequently, classifying phone usage sessions through a threshold delimiter results in many false positives.

Given our findings, we believe that solely relying on the use of phone standby time analysis to classify smartphone usage gaps is not reliable. However, if a researcher insists on using a constant arbitrary threshold, we encourage fellow researchers to consider using a T=45 seconds threshold (Figure 8) when analysing smartphone usage data. This threshold value minimises the error based on our analysis. However, it still performs rather poorly in absolute terms (68% accuracy) and, given the relative low number of 17 participants, is not generalizable. Constant Classifiers do not generalise well; therefore, we should expect a low accuracy in this context. Constant Classifiers use only one feature for classification, which results in a high bias. On the contrary, Similar Sets and One Rule Classifiers can potentially generalize because they use more data for classification in order to infer whether the class is a continuous session or a new session. However, in the context of our study they performed poorly as well. This means that using a T threshold is actually preferable for classification. In addition, individual user modelling would enable further improvements to the model. As visible from Table 2, large differences exist between users on the ratio of continuous versus new usage sessions.

Modelling intermittent smartphone use

Our findings form a basis for revisiting the revisitation analysis method [7] in the context of smartphone usage in general. Recent work has conducted revisitation analysis of individual smartphone *application* use [19], mirroring earlier studies that looked at revisitation patterns for web browsing on desktops [1] and smartphones [32]. While researchers use most of this work to characterise smartphone applications and websites, some researchers have used similar methods to profile desktop users [28] or smartphone users [19]. The majority of prior work employing revisitation analysis has relied on *post-hoc* usage traces and has not included participant-labelled data. Revisitation analysis has yet to consider the motives and intents of users. These can be very relevant, as one participant noted: "*Sometimes I check my phone really frequently and sometimes I forget it somewhere and only check it in the end of the day.*" (P02). Here we demonstrate that, in the context of smartphone usage sessions, the exponential binning time-threshold values used in revisitation analysis [19] is not adequate to characterise the purpose of a revisit (*i.e.*, returning to use smartphone).

However, to address this limitation, one can slightly adjust the bins used in revisitation analysis by considering the results shown in Figure 7, *i.e.*, the number of consecutive *continuous* sessions according to participants' labelled data. We found that the majority of phone usage sessions that contain a gap (*i.e.*, phone went to standby mode) consist of only one additional continuous session. The frequency of instances with a higher number of cumulative continuous sessions quickly declines as the number of continuous sessions rises. By also considering the overall ratio between new and continuous sessions, we can conclude that in most cases a person will complete their objective within a single usage session, or within two consecutive sessions. Thus, a revisitation analysis of smartphone usage can leverage binning time values by adopting our T threshold value for the first bin (45 seconds). This can more accurately profile users based on their intermittency of use, since we show that users do not get that often interrupted on their smartphone and tend to complete their objectives in one or two "visits" to their device.

Understanding gaps to improve smartphone interaction

Our work studies the gaps in smartphone interaction, and through user-labelling we are able to study whether they are interruptions to users' objectives. Previous research has typically made assumptions/speculations about the

meanings of gaps, mostly because researchers have primarily focused on the application aspect of analysis. For example, by understanding how people use their phone in a sporadic manner it is possible to design interfaces and technology that supports intermittent application usage [3,12] or provide visual cues for incomplete tasks [12,22]. Furthermore, with the increasing prevalence of smartwatches and other wearable devices, it is increasingly interesting to understand intermittent application use across such devices [2,34].

Similarly, Pielot *et al.* [26] have shown that phone users check most notifications within a short period after arrival, even if phones are in silent mode. However, their results lack the insight about whether users *unlock* their phone after notifications emerge. The results shown in Table 3 indicate that user habits are highly divergent. Our interviews also confirm this divergence. Some participants may have reacted to notifications as described when asked about their phone usage habits: "*When I received notifications and when I need to use the phone (like search).*" (P8), while others seem to have a checking schedule: "*I take short peeks at it every half an hour. Duolingo once per day, [for] about 10-15 mins.*" (P07). A number of participants tended to passively wait for a notification before they unlock their phones. Thus, a threshold value for T may not help to determine whether this type of user start a *new* usage session or not.

To enable researchers to consistently report their findings, we need consistent definitions and metrics that are comparable across studies, experiments, and devices [6]. We argue that consistency is lacking in the definitions found in literature, leading to widely differing results. For instance, reports from *mean* session length range from 65 seconds [24] to 4:42 minutes (unlocked device usage) [16]. In our study, we report the mean duration of usage sessions to be 4:43 minutes, placing our findings at the upper bound of prior results.

There are three factors that can help explain this discrepancy. First, our calculation is unique in the sense that it considers the user-labelled data. This allows us to employ the definition of a usage session focused on the *actual* completion of objectives suggested in [18], as opposed to the observation of a usage session as a fixed time window. With user-labelled data, we calculate the duration of a phone usage session from a task-completion perspective, and thus consider *continuous* sessions to extend their predecessor. As a result, we obtain a longer mean session time of almost 5 minutes in length.

Second, some researchers do not incorporate 'idle usage' and device timeout values and subtract these from their recorded phone usage session. We did not perform such subtractions in our calculations, as they are an integral part of mobile phone usage. Third, since participants have to respond to our ESM question every time they unlocked their phone, the duration of the usage session was slightly increased – we required our participants to answer a question prior to their actual phone usage. However, this extension is short, and is typically in the range of a few seconds (median of 2 seconds).

It is also interesting to note that during the debriefing, most participants felt that the experiment was unobtrusive, and therefore did not substantially change their behaviour or use of the phone. One participant likened the ESM to a screen lock: "*No, it did not affect my regular phone usage. It was just normal. Similar to a screen lock.*" (P05). Another participant mentioned that he only felt a small change during the first day until he got used to it: "*Not really, during the day 1 I was a bit slower at answering my phone.*" (P06).

Lessons learned

Our analysis shows that the prediction accuracy achieved in our classification of smartphone usage gaps is relatively low. However, for the data in our sample, still results in a higher accuracy than methods currently applied in the literature (*e.g.*, arbitrary threshold of 30 seconds). This low accuracy does not only demonstrate a weakness in the current literature, but also limits the potential of this work to contribute to the design of future services and applications for mobile device users.

A more reliable identification of smartphone usage gaps has the potential to improve the user experience for end-users through smarter applications and services. For example, the content provided to the user upon unlocking a device can depend on the results of the classifier. Devices could achieve this by showing the user information previously interacted with, or returning to an overview of available applications or services. Furthermore, this allows the operating system to infer what information to retain in working memory, thereby decreasing required system resources. We also consider knowledge on usage gaps to be valuable for the design of more proactive services that inform the user based on the users' context.

Limitations

The work presented in this paper has several limitations. Because of the study's reliance on user-labelled ESM data on phone unlock, it is not possible to collect data of phone sessions where the screen is active but locked (state "Locked session" in Figure 1). This can, for example, include the glancing of time and notifications, as discussed in [16], or access to certain functionality through the notification drawer (*e.g.*, music controls). However, these issues did not surface during post-study interviews and we therefore expect them to have only a marginal effect on the study results.

In addition, we realise that a user can in fact pursue multiple objectives during a phone usage session, or indeed during a single application session [3]. Hence, the concept of session does not precisely align with objectives. However, our analysis only considered whether objectives

could span multiple usage session, and to this end, our analysis serves its purpose well.

CONCLUSION

Our work has provided a systematic model of smartphone usage along with definitions of what phenomena and behaviour researchers can study. We subsequently investigate an important assumption present in literature: that researchers should ignore brief gaps in interaction. Previous work has used a range of arbitrary time thresholds for identifying "brief" gaps, which has resulted in incomparable findings across studies. Our work shows that the use of such a threshold is problematic and leads to error in general. However, researchers can minimise the error by setting the threshold to 45 seconds as opposed to an arbitrary value. We also find that, perhaps surprisingly, classifiers are not able to outperform the use of constant thresholds. Other researchers can readily adopt our findings to inform their work. Future work could expand this work by actively predicting the time gap between two sessions.

ACKNOWLEDGEMENTS

This work is partially funded by the Academy of Finland (Grants 276786-AWARE, 285062-iCYCLE, 286386-CPDSS, 285459-iSCIENCE), and the European Commission (Grants PCIG11-GA-2012-322138 and 645706-GRAGE).

REFERENCES

1. Eytan Adar, Jaime Teevan and Susan T. Dumais. 2008. Large Scale Analysis of Web Revisitation Patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1197-1206. http://dx.doi.org/10.1145/1357054.1357241
2. Eirik Årsand, Miroslav Muzny, Meghan Bradway, Jan Muzik and Gunnar Hartvigsen. 2015. Performance of the First Combined Smartwatch and Smartphone Diabetes Diary Application Study. *Journal of Diabetes Science and Technology* 9, 3: 556-563. http://dx.doi.org/10.1177/1932296814567708
3. Nikola Banovic, Christina Brant, Jennifer Mankoff and Anind Dey. 2014. ProactiveTasks: The Short of Mobile Device Use Sessions. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*, ACM, 243-252. http://dx.doi.org/10.1145/2628363.2628380
4. Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger and Gernot Bauer. 2011. Falling Asleep with Angry Birds, Facebook and Kindle: A Large Scale Study on Mobile Application Usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, ACM, 47-56. http://dx.doi.org/10.1145/2037373.2037383
5. Juan P. Carrascal and Karen Church. 2015. An In-Situ Study of Mobile App & Mobile Search Interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 2739-2748. http://dx.doi.org/10.1145/2702123.2702486
6. Karen Church, Denzil Ferreira, Nikola Banovic and Kent Lyons. 2015. Understanding the Challenges of Mobile Phone Usage Data. In *International Conference on Human-Computer Interaction with Mobile Devices and Services*, 505-514. http://dx.doi.org/10.1145/2785830.2785891
7. Andy Cockburn and Bruce McKenzi. 2001. What Do Web Users Do? An Empirical Analysis of Web Use. *International Journal of Human-Computer Studies* 54, 6: 903-922. http://dx.doi.org/10.1006/ijhc.2001.0459
8. Sunny Consolvo and Miriam Walker. 2003. Using the Experience Sampling Method to Evaluate Ubicomp Applications. *IEEE Pervasive Computing* 2, 2: 24-31. http://dx.doi.org/10.1109/MPRV.2003.1203750
9. Tilman Dingler and Martin Pielot. 2015. I'll Be There for You: Quantifying Attentiveness Towards Mobile Messaging. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, 1-5. http://dx.doi.org/10.1145/2785830.2785840
10. Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan and Deborah Estrin. 2010. Diversity in Smartphone Usage. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, ACM, 179-194. http://dx.doi.org/10.1145/1814433.1814453
11. Denzil Ferreira, Eija Ferreira, Jorge Goncalves, Vassilis Kostakos and Anind K. Dey. 2013. Revisiting Human-Battery Interaction with an Interactive Battery Interface. In *International Joint Conference on Pervasive and Ubiquitous Computing* , ACM, 563-572. http://dx.doi.org/10.1145/2493432.2493465
12. Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, Louise Barkhuus and Anind K. Dey. 2014. Contextual Experience Sampling of Mobile Application Micro-Usage. In *International Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, 91-100. http://dx.doi.org/10.1145/2628363.2628367
13. Denzil Ferreira, Vassilis Kostakos and Anind K. Dey. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT* 2, 6: 1-9. http://dx.doi.org/10.3389/fict.2015.00006
14. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian I. H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1: 10-18. http://dx.doi.org/10.1145/1656274.1656278
15. Joel M. Hektner, Jennifer A. Schmidt and Mihaly Csikszentmihalyi. 2007. *Experience Sampling Method: Measuring the Quality of Everyday Life.* Sage.

16. Daniel Hintze, Rainhard D. Findling, Muhammad Muaaz, Sebastian Scholz and René Mayrhofer. 2014. Diversity in Locked and Unlocked Mobile Device Usage. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ACM, 379-384. http://dx.doi.org/10.1145/2638728.2641697

17. Robert C. Holte. 1993. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning* 11, 1: 63-90. http://dx.doi.org/10.1023/A:1022631118932

18. Bernard J. Jansen, Amanda Spink, Chris Blakely and Sherry Koshman. 2007. Defining a session on Web search engines. *Journal of the American Society for Information Science and Technology* 58, 6: 862-871. http://dx.doi.org/10.1002/asi.20564

19. Simon L. Jones, Denzil Ferreira, Simo Hosio, Jorge Goncalves and Vassilis Kostakos. 2015. Revisitation analysis of smartphone app use. In *International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 1197-1208. http://dx.doi.org/10.1145/2750858.2807542

20. Rosie Jones and Kristina L. Klinkner. 2008. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ACM, 699-708. http://dx.doi.org/10.1145/1458082.1458176

21. Reed Larson and Mihaly Csikszentmihalyi. 1983. The Experience Sampling Method. In *Flow and the Foundations of Positive Psychology* (eds.). Wiley Jossey-Bass, San Francisco, 15, 41-56.

22. Luis A. Leiva, Matthias Böhmer, Sven Gehring and Antonio Krüger. 2012. Back to the app: the costs of mobile application interruptions. In *MobileHCI'12*, 291-294. http://dx.doi.org/10.1145/2371574.2371617

23. Terence R. Mitchell, Leigh Thompson, Erika Peterson and Randy Cronk. 1997. Temporal Adjustments in the Evaluation of Events: The "Rosy View". *Journal of Experimental Social Psychology* 33, 4: 421-448. http://dx.doi.org/10.1006/jesp.1997.1333

24. Earl Oliver. 2010. The challenges in large-scale smartphone user studies. In *Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement - HotPlanet '10*, Art. 5. http://dx.doi.org/10.1145/1834616.1834623

25. Antti Oulasvirta, Tye Rattenbury, Lingyi Ma and Eeva Raita. 2012. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing* 16, 1: 105-114. http://dx.doi.org/10.1007/s00779-011-0412-2

26. Martin Pielot, Karen Church and Rodrigo de Oliveira. 2014. An In-situ Study of Mobile Phone Notifications. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*, ACM, 233-242. http://dx.doi.org/10.1145/2628363.2628364

27. Martin Pielot, Tilman Dingler, Jose S. Pedro and Nuria Oliver. 2015. When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 825-836. http://dx.doi.org/10.1145/2750858.2804252

28. Philipp Pushnyakov and Gleb Gusev. 2014. User Profiles Based on Revisitation Times. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, 359-360. http://dx.doi.org/10.1145/2567948.2577380

29. Ahmad Rahmati and Lin Zhong. 2010. A Longitudinal Study of Non-Voice Mobile Phone Usage by Teens from an Underserved Urban Community. *Computing Research Repository*.

30. Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber and Albrecht Schmidt. 2014. Large-scale Assessment of Mobile Notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 3055-3064. http://dx.doi.org/10.1145/2556288.2557189

31. Tapio Soikkeli, Juuso Karikoski and Heikki Hämmäinen. 2011. Diversity and End User Context in Smartphone Usage Sessions. In *International Conference on Next Generation Mobile Applications, Services and Technologies*, IEEE, 7-12. http://dx.doi.org/10.1109/NGMAST.2011.12

32. Chad Tossell, Philip Kortum, Ahmad Rahmati, Clayton Shepard and Lin Zhong. 2012. Characterizing Web Use on Smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2769-2778. http://dx.doi.org/10.1145/2207676.2208676

33. Niels van Berkel, Chu Luo, Denzil Ferreira, Jorge Goncalves and Vassilis Kostakos. 2015. The Curse of Quantified-Self: An Endless Quest for Answers. In *Adjunct Proceedings of International Joint Conference on Pervasive and Ubiquitous Computing Adjunct*, 973-978. http://dx.doi.org/10.1145/2800835.2800946

34. Robert Xiao, Gierad Laput and Chris Harrison. 2014. Expanding the Input Expressivity of Smartwatches with Mechanical Pan, Twist, Tilt and Click. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 193-196. http://dx.doi.org/10.1145/2556288.2557017

35. Tingxin Yan, David Chu, Deepak Ganesan, Aman Kansal and Jie Liu. 2012. Fast app launching for mobile devices using predictive user context. In *MobiSys*, 113-126. http://dx.doi.org/10.1145/2307636.2307648