OpenMPower: An Open and Accessible Database About Real World Mobile Devices

A. Corna, A. Damiani, M. Ferroni, A. A. Nacci, D. Sciuto, M. D. Santambrogio
{andrea.corna, andrea3.damiani}@mail.polimi.it
{matteo.ferroni, alessandro.nacci, donatella.sciuto, marco.santambrogio}@polimi.it
Dipartimento di Elettronica Informazione e Bioingegneria
Politecnico di Milano
Milano, Italy

***Abstract*—In the last decade we have witnessed the birth and dramatic growth of mobile devices, from cellular- to smart-phones. Despite the huge amount of information achievable from an always-connected reality, researchers that work in the mobile devices field fight against the impossibility to explore, inspect and test their work on such a vast set of possible environments, use case scenarios, hardware and software platforms the smart mobile world is composed of. This pushed the need of a wide open dataset of real world data coming from devices in their real usage context, properly anonymized and conveniently organized to be searchable and accessible. In this paper, we present a platform that brings such a dataset to researchers of the next generation of mobile devices.**

***Keywords*—opendata, database, mobile devices, activity, logging, smartphones**

I. Introduction

In the last decade we have witnessed the birth and dramatic growth of the smartness of mobile devices [9]. This process began with mobile connectivity becoming the core facility in cellular phones [2], allowing more intelligence to develop on board of these devices [14].

Nowadays the increasing demand for smarter and context-aware mobile devices [9], is applying a growing pressure on researchers and developers in order to solve the issues that currently affect these pieces of technology and make new visions a reality. This goal is hard to reach because of the impossibility to explore, inspect and test the works about the smart mobile environment, especially because this world is a vast, chaotic collection of ever-mutating settings, targets and configurations [16]. Moreover, many early research studies about achieving better battery life, performance, security, etc. for mobiles are now proving to be inadequate; [28] and [8] firmly highlighted the need for new methodologies able to better keep up with this evolutionary trends. Techniques and algorithms that adapt the better to each current specific hardware, software, configuration and user are now the path to be unleashed and thoroughly analyzed. The main issue is that every researcher should validate his/her methodology on a number of cases that may easily become overwhelming. For instance, a study like the one of N. Vallina-Rodriguez et al. in [29] requires a great effort in generating a number of models to highlight the impact of the users behaviors on mobile devices. This leads to the necessity of a wide open dataset that gathers all the information that is important to support multiple researches on the mobile world; a dataset that records real world data coming from the users' real devices, properly anonymized and protected against potential privacy invasions, organized to be accessible to all interested players. In this paper we present a platform that integrates and supports such a dataset — featuring more than 60 million complete records corresponding to samples of the state of almost 5 hundreds of Android devices in the real world, obtained in a time window of the last 3 years (see Figure 1)— and that is available to all the people from the academic and research domains, involved in making the mobile world better and better.

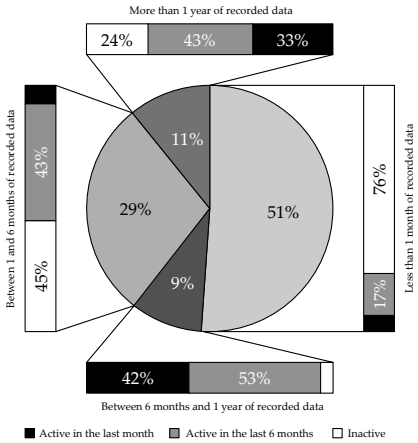

Fig. 1. Devices Longevity (457 devices)

We believe that the huge quantity of structured data we are releasing is surely useful for the research community, even if this paper may not present a direct technical contribution *per se*. We hope that this dataset could become a shared and common base to support, compare and reproduce effectively all the main works in the field of mobile devices.

II. The importance of being mobile

There is an increasing number of research areas about mobile devices which is gaining critical importance in guiding the evolution of technology in mobility [28]. This leads inevitably to discover and target a growing number of issues. The most evident one is about power consumption and battery life. Since research on new batteries seems not to keep up with always increasing power requirements, mainly three approaches have been developed to build power models

to extend devices battery life [8]: off-line modeling, that proved not to scale with the growth and diversification of the mobility market; on-line modeling, with dedicated hardware and software allowing a better description of real cases [29]; machine learning enhanced modeling, to better tackle the most important variable in power consumption equation - the user behavior [5]. Another important issue rises from the fact that smartphones are now widely used to carry out an always increasing number of daily tasks (electronic payments, identification, health monitoring, etc.) and so they require high level of dependability and security [30]. In the latter field, many studies have been carried out in order to overcome all the limitations of signature-based malware detection, e.g. L. Liu *et al.* in [15] effectively developed an app that detects malware infections on mobile devices by analyzing anomalous fluctuations in power consumption.

Taking a closer look to all the studies that address all the issues we listed before, a common caveat can be identified: they all use a small and non-representative set of devices to both support and validate their approach. For example, N. Banerjee *et al.* carried out an influential study in [1] highlighting the importance of using on-line device state traces in correctly characterizing the mobile power consumption. However, this study bases all its important results on the examination of only 10 identical PDAs. Also H. Falaky et al. in [4] state: *"A key limitation of our work is the small user populations of our datasets"*, stressing unequivocally the importance of having access to a big and varied dataset in order to draw significant and useful conclusion on experimental results. Given the necessity of utilizing such a dataset for almost any study in the mobile technology field, many researchers began to develop specific logging app and to distribute them on the main stores in order to try to reach the vastest public possible. Apps such as LiveLab ([26]), MyExperience ([11]) and MPower ([7]) differently interpreted the need of an on-line logging facility, leading to duplicate a large part of the data collected due to the fact that each dataset was crafted specifically as a means to support the main goal of each app and so was only accessible by the project it belonged to.

III. Open Data: share what you know

Data is fundamental to provide supporting evidence for the publication corpus of any kind of scientific knowledge, and public availability ensures greater level of transparency and reproducibility of any study. This trend is very common in biology and medicine fields, in which the large availability of data is faced with the scarcity of computational resources. Moreover, this emerging methodology for data spreading, is stimulating the interest of a very diverse set of players: from commercial corporations, such as Apple, which recently disclosed its *ResearchKit* to support medical data sharing [17], to eminent research centers, such as CERN, which announced the will to make public the entire data sets related to collisions experiments [18]. In particular, the open data concept is based on the idea that *"the resource data should be openly available to the maximum extent possible"* [24]. The goals of the open data movement are similar to those of other "open" movements such as open source, open hardware, open content, and open access. To tackle all the criticalities in sharing data, the accesses are managed by different organizations, either private or public, that define the rules and the characteristics that a sharing information system must have. An example is the *Open Knowledge Foundation* [21]. Considering all these characteristics, our goal is to propose a system that is able to share information in the best way possible.

IV. Logger application

In the last three years, we have been collecting a constantly increasing amount of data about mobile devices "into the wild", i.e., in their real usage context, thanks to our experience with the MPower project. The system is composed of two main components: a logging application on the mobile device that gathers its "state" and a remote cloud server infrastructure for data analysis. The logging application is built on top of the standard Android SDK and it is currently available on the Google Play Store [23], no device rooting is required and it is compatible with almost any Android device. This application has been developed with a strong requirement in mind: the logging operations should influence as little as possible the device power consumption and its standard behavior. In order to introduce no power consumption due to data transmission, information is sent to our servers only when the device is plugged into an external power source and connected to a WiFi network, to avoid unnecessary 3G/4G data transfer. Then, we put a significant effort in making the data collecting phase the most nonintrusive possible from an energy point of view: we came out with the implementation of an highly optimized *Sense Library*, i.e., a set of Java classes that encapsulates all the logic needed to support different Android versions and that follows the Android best practices for an effective, efficient and maintainable source code. The hardware features that can be accessed by means of the *Sense Library* and that are included in the device state we sample are: screen (e.g., ON/OFF state, brightness level, orientation, etc.), battery (e.g., level of charge, temperature, etc.), CPUs (e.g., current frequency, usage, etc.), Audio (e.g., ON/OFF state, level, etc.), Mobile data (e.g., ON/OFF state, received and transmitted bytes, data type, etc.), Wi-Fi (e.g., ON/OFF state, received and transmitted bytes, signal strength, etc.), Bluetooth (e.g., ON/OFF state, etc.) and GPS (e.g., ON/OFF state) connections, as well as the foreground and the background applications currently running in a certain time instant. The source code of this library is released under the GNU LGPL license (details and technical report at [23]). Finally, the sampling frequency of the monitoring activity was the last crucial aspect to keep into consideration: on the one hand, a high frequency allows accurate profiling of device's status, while, on the other hand, a frequent activity may imply a significant overhead on the device power consumption. After several empirical tests, we decided to log the device state every 10 seconds, as also suggested by the work presented in [31]. Moreover, our logging application does not make use of any *wakelock* , as this could potentially lead to unnecessary energy consumption: as a consequence, the Android system may fall into a *deep sleep* mode, thus implying discontinuous registrations of the system state. The logging application, as well as all the other Android

services, is resumed as the system leaves this low power mode [22]. With the proposed approach, we guaranteed a reasonable compromise of the three characteristics of compatibility, energy efficiency and adaptability, by targeting all the Android devices, showing that the logging application interferes in a negligible way on the power consumption, still being able to gather mobile systems information in real usage context.

V. Accessing our open dataset

We hereby propose an open platform that is specifically thought to answer the need the research community is expressing: a common open dataset about mobile technology, sampled in a real world environment and on a large set of distinct devices. We decided to disclose to every interested researcher the dataset we built to support the MPower project, taking care of all the issues about anonymization and privacy protection: the unique identifier of the specific mobile device is anonymized in the sanitization process of copying those data to the openMPower database, and no other sensitive information about the user (e.g., his/her GPS location) is stored. In Table I we report the dataset general characteristics, while in Figure 1 we give representative breakdowns in terms of the longevity and activity of the devices, and, finally, in Figure 2, we broadly analyze its market coverage. It is interesting to highlight that the size of the dataset, even if it may seem limited with respect to the mobile market size and anyway constantly growing in both number of registered devices and duration of traces, is already more than sufficient to replicate a wide range of studies carried out about mobile devices in the past. To provide an exemplification of the coverage of our dataset, without any claim of completeness, we examined the papers presented in a survey on mobile energy management by Vallina-Rodriguez *et al.* in [28]: only two studies took advantage of a broader number of devices than the one available in openMPower ([27] and [13]), the first one is agnostic of the devices characteristics and focuses only on the user interaction patterns with Apps while the second one was carried out in a simulated environment (3G connectivity simulated via Wi-Fi tuning) – both out of the scope of openMPower that focuses on making data from real usage of devices. The other studies surveyed in [28] can be divided into two categories, the ones connected to laboratory controlled environments of experimentation – again, out of the scope of openMPower – and the ones connected with real world usages. In the vast majority of the cases these studies involved a very limited number of devices and a small number of data features with respect to what is currently available in our dataset. This allows us to state that openMPower allows to replicate those studies, without requiring each research team to develop its own logging facilities in-house. Finally, in most of the studies we analyzed, the logging methodologies and applications are not fully reported, rising problems at the moment of comparison; our dataset instead removes these kind of issues between the studies it supports. For completeness, we must note that the only studies in [28] that could not be completely replicated leveraging our dataset are the one related with the precise geographical location of the users, because of our restricting privacy policy.

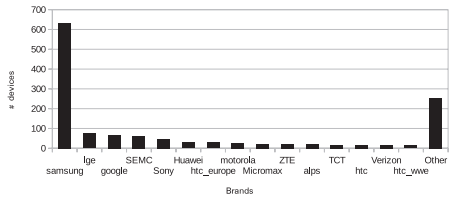

Fig. 2. Brand Distribution (all registered devices)

Our goal is to offer a single central point from where all interested researchers can obtain real and significant data on mobile devices behavior. This will firstly avoid them to develop and deploy their own logger app and focus only on the research objective; secondly, it will constitute a base for comparing different studies, taking out of the equation all the discrepancies among data collection methods and sample compositions. Our platform can be framed into the open data movement and complements works such as [19] focusing more on the mobile device itself than on the pure user behavior.

TABLE I. Dataset Characteristics

Characteristic	**Fields**
Dataset dimension	120 GB
Number of brands	139
Number of models	650
Number of devices	706[1]
Data collection period	3 years

Providing these data, we enable other researchers to make their own contribution to this topic even if they don't have proper infrastructures to collect data or even if they don't have data at all. For this reason, we have taken into account all general characteristics of a system that wants to share data: first simplicity of use, then as open as possible, paying attention to users' privacy and finally possibility to manipulate and reuse a copy of data in order to do some research projects. The project dataset is made available under the Open Database License[20].

VI. The platform architecture

The application is composed by two Python applications developed using the web framework Flask: the first provides some RESTFul API in order to access data, while the second one is a control panel for user management. The results are provided using the JSON data representation, so that they can be easily used and manipulated by clients.

A. API Server

The first application implements the RESTful paradigm in order to give information to the clients. In particular, all possible operations are GET requests in which the client has to specify the token received during the registration phase: when a request comes to the application, the token is controlled to ensure that it is still valid and active, and all information related to that url are returned. We give different APIs in order to answer to various users' requests. The API structure is defined and detailed with all data categories available at *http://openmpower.necst.it/documentation*. Additional materials about the platform usage can be found at *http://mpower.necst.it/opendata*. All queries related to a

[1] 65% of which (457) with at least 1 month of traces

single device data are returned in descendant order using the timestamp field and every request consists of 1000 records at most. It is possible to specify the parameter *page* in the request in order to select a different portion of dataset.

B. Site application

The site application provides the functionality to manage the user requests: in particular, it provides a user-friendly interface to ask for application access, check the validity of his token, communicate with the administrator of the system and ask for more access tokens. Moreover we have inserted a section in which we propose a possible use of data, as described in subsection VII-A. The site can be reached at *http://mpower.necst.it/opendata*.

VII. Case Studies

In previous sections we underlined the importance and the possibilities coming from the diffusion of mobile devices. In this section we propose some proof of concepts, to show how our dataset can be handled on those open problems. It is important to underline the fact that these examples are deliberately far from being complete studies in which we want to state conclusions, since their intended focus resides only in showing the potentials of the system we propose and its capability to cover different areas of interest.

A. Impact of antivirus applications

A lot of users installed on their devices an AntiVirus System (AVS) application to protect it from malwares. AVSes are said not to impact significantly on battery life. The only way to prove this claim is to look for statistical evidence on a real dataset, like the one we propose. The idea is to evaluate the impact of AVS applications considering the mean Time-To-Live [7] of devices. In particular, we want to compare the behavior of devices in which we have some data with no AVS installed and some other data with an AVS installed. The analysis was done considering different device models; information was compared only if two different devices of the same model respect the conditions said before. In general, the expectation on the results was that a device with an AVS application installed should have had a lower TTL, since a constant monitoring service may require a non-negligible amount of energy. However the outcome, shown in Figure 3, is a bit different; in fact, there is not any strong evidence of the hypothesis mentioned above, due to the fact that a significant portion of results confute it.

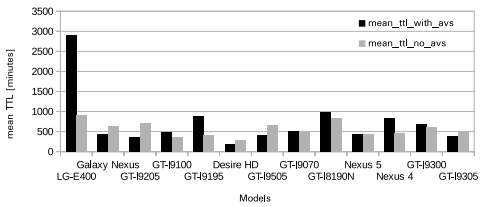

Fig. 3. Analysis of the impact of AVSes on our dataset

B. Battery emergency

Nowadays we can do a lot of activities with our smartphones, which have become the most powerful instrument we may use to interact with the World. Since all these actions accelerate and influence the battery discharge, we are often forced to give up using our devices in order to preserve the remaining charge. An interesting point is to determine how often a user is in a critical situation and in which time intervals. Again, in the last few years researchers have put a great emphasis on this topic, conducting very large scale experiments [6], in order to extract valuable patterns to support users in preventing "battery emergencies". With the dataset we provide, a researcher can extract these information for every user. In Figure 4 we report the percentage distribution over the day of the times in which the battery life was less than 10%. One consideration we can easily draw is that many users experience low batteries way before evening time.

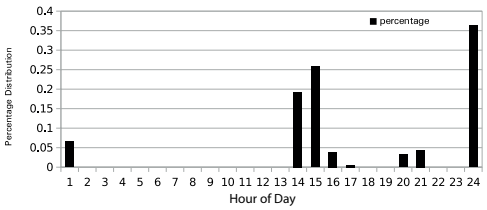

Fig. 4. Low Battery Distribution Over The Day

C. YouTube Bandwidth consumption

Current mobile devices have all the necessary resources and possibilities to play multimedia video contents, taking advantage of application stream services such as YouTube. This type of application has to take into account different issues: first a good quality of streaming in an acceptable time, secondly a low impact on data usage and energy consumption. These aspects are very useful and have been object of a lot of studies such as [25], [10], [3] and [12]. In particular, a very interesting information could be the isolation and characterization of the average consumption of stream data provided by YouTube application in Android devices daily usage. More in details, we want to see if there are some implementations of economy policy in order to reduce data mobile usage in case of using a 3G connection. The results of such analysis are reported in Figure 5: first of all we compare the mean usage in the case of a wifi communication, respect to one with 3G network; secondly, we compare the results of these two approaches with the general consumption.

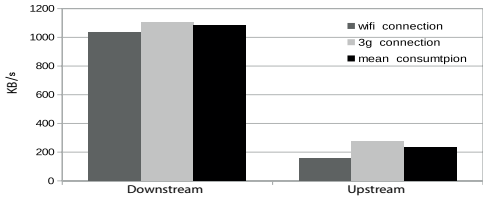

Fig. 5. YouTube Average Data Traffic

VIII. Limitations and Lesson Learned

The system we provide has some limitations related to poor hardware performances. For this reasons we have decided to limit the size of information returned by every query to a maximum of 1000 records, and to require the users to register in our system. With respect to the web site, it returns a warning message about the security of the connection: it is due to the

fact that our SSL certificates are not signed by a certification authority. Independently of these problems, with this work we have learned how to develop a system that is able to openly share data, considering all the preeminent aspects related to this topic. Our intent is to provide a simple interface that users can leverage in order to retrieve information about mobile devices behavior in the real world. Putting the privacy of the users providing data first, we understood how important is to precisely protect all sensible information from malicious accesses and misuse, balancing the need to openly share this information and preserve the anonymity of the sources. For these reasons we adopted some techniques in order to limit the operations that can be performed on data and to verify the users' identity. Moreover we gained an important insight on the importance of sharing information: many previous researches mentioned in this paper could have been realized with much less effort using openMPower and now can be reproduced with more and real data allowing the entire scientific community to verify, validate and compare the results obtained.

IX. Conclusions

In this work we presented our open platform to access a dataset containing real life information on mobile devices, collected in everyday conditions. Moreover, our dataset will continue to expand harvesting always new data in time from both already registered devices and others that will join the program.

Our platform fills the need of a variegated, vast, reliable and always up-to-date database that researchers require in order to carry out works that can be considered valid in real scenario. In fact, researchers tend to limit their studies to one or few more devices, to laboratory controlled environments and to non-realistic usage patterns because of the great burden that is associated to retrieve, within the time window of a single study, the great amount of data necessary to solidly validate a work. After having requested an authentication token to the administrators, everyone will get read access to the platform and can query it via RESTful APIs. We showed that many different topics can be investigated relying on our open data platform, thanks to the data neutrality principle it founds on, by simply imposing different cuts to the returned data themselves. Moreover, a wide use of our platform in different works will simplify their comparison thanks to the absence of reciprocal differences in data collection and processing algorithms that may lead in biased comparative analysis. Future developments of our platform will include the activation of write APIs that will allow researchers to give a direct contribution in extending the open dataset.

References

[1] N. Banerjee, A. Rahmati, M. D. Corner, S. Rollins, and L. Zhong. *Users and batteries: Interactions and adaptive energy management in mobile systems*. Springer, 2007.

[2] S. J. Barnes and S. L. Huff. Rising sun: Imode and the wireless internet. *Commun. ACM*, 2003.

[3] J. Erman, A. Gerber, K. Ramadrishnan, S. Sen, and O. Spatscheck. Over the top video: the gorilla in cellular networks. ACM, 2011.

[4] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin. A first look at traffic on smartphones. ACM, 2010.

[5] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in smartphone usage. ACM, 2010.

[6] D. Ferreira, A. K. Dey, and V. Kostakos. Understanding human-smartphone concerns: a study of battery life. Springer, 2011.

[7] M. Ferroni, A. Cazzola, D. Matteo, A. A. Nacci, D. Sciuto, and M. D. Santambrogio. Mpower: gain back your android battery life! ACM, 2013.

[8] M. Ferroni, A. Cazzola, F. Trovo, D. Sciuto, and M. D. Santambrogio. On power and energy consumption modeling for smart mobile devices. IEEE, 2014.

[9] M. Ferroni, A. Damiani, A. A. Nacci, D. Sciuto, and M. D. Santambrogio. coda: An open-source framework to easily design context-aware android apps. IEEE, 2014.

[10] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao. Youtube everywhere: Impact of device and infrastructure synergies on user experience. ACM, 2011.

[11] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones. ACM, 2007.

[12] M. A. Hoque, M. Siekkinen, J. K. Nurminen, and M. Aalto. Dissecting mobile video services: An energy consumption perspective. IEEE, 2013.

[13] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl. Anatomizing application performance differences on smartphones. ACM, 2010.

[14] M. Ilyas and S. A. Ahson. *Smartphones*. Intl. Engineering Consortiu, 2006.

[15] L. Liu, G. Yan, X. Zhang, and S. Chen. Virusmeter: Preventing your cellphone from spies. Springer, 2009.

[16] C. Mascolo. The power of mobile computing in a social era. *IEEE internet computing*, 2010.

[17] Apple Inc. Apple introduces researchkit, giving medical researchers the tools to revolutionize medical studies. *https://goo.gl/XksYF4*.

[18] CERN. Cern makes public first data of lhc experiments. *http://goo.gl/fvHDr1*.

[19] MIT Human Dynamics Lab. Reality commons. *http://goo.gl/8PXfKO*.

[20] Open Knowledge Foundation. Open database license. *http://goo.gl/B2Hh4z*.

[21] Open Knowledge Foundation. Open knowledge. *https://okfn.org/*.

[22] R. Meier. *Professional Android 4 Application Development*. Wrox, 3rd edition, May 2012.

[23] NECST-Laboratory. Mpower project. *http://goo.gl/R3F3YW*.

[24] P. P. A. Others. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 2004.

[25] A. Rao, A. Legout, Y.-s. Lim, D. Towsley, C. Barakat, and W. Dabbous. Network characteristics of video streaming traffic. ACM, 2011.

[26] C. Shepard, A. Rahmati, C. Tossell, L. Zhong, and P. Kortum. Livelab: measuring wireless networks and smartphone users in the field. 2011.

[27] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring serendipity: connecting people, locations and interests in a mobile 3g network. ACM, 2009.

[28] N. Vallina-Rodriguez and J. Crowcroft. Energy management techniques in modern mobile handsets. *Communications Surveys & Tutorials, IEEE*, 2013.

[29] N. Vallina-Rodriguez, P. Hui, J. Crowcroft, and A. Rice. Exhausting battery statistics: understanding the energy demands on mobile handsets. ACM, 2010.

[30] J. Wright, M. E. Dawson Jr, and M. Omar. Cyber security and mobile threats: The need for antivirus applications for smart phones. *Journal of Information Systems Technology and Planning*, 2012.

[31] L. Zhang, B. Tiwana, Z. Qian, Z. Wang, R. P. Dick, Z. M. Mao, and L. Yang. Accurate online power estimation and automatic battery behavior based power model generation for smartphones. In *Proceedings of the eighth IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis*. ACM, 2010.