Device Analyzer

Daniel T. Wagner, Andrew Rice, Alastair R. Beresford
University of Cambridge Computer Laboratory

New hardware capabilities and integrated application market stores have created complex usage patterns in modern smart phones. Understanding these patterns is important to both smart phone providers and their customers. Hardware manufacturers might make use of this data when considering design trade-offs of functionality against cost and form-factor. Summarised data of handset use is also useful for device owners trying to stay within data limits on mobile networks or even to select the best tariff. The idea of *personal analytics* takes this to an extreme in which one collects detailed statistics of many life aspects.

Our Device Analyzer application aims to collect a large-scale research data-set of phone usage. Device Analyzer gathers data about running processes, wireless connectivity, the phone's location, GSM communication, battery state and a number of system parameters. Our work is inspired by studies such as MIT Reality Mining which tracked 100 students over the course of the 2004–2005 academic year [1]. We want to extend the range of data collected as well as the breadth of the targeted population. Today's smart phones are able to capture much richer data than was available some years ago while at the same time the Android Market allows us to distribute the app to a broader audience than was possible before.

Data Collection Device Analyzer makes use of events from the Android operating system to record many changes in system state. However, events are not available for some types of data, such as process information or network traffic, and so these are gathered by polling. An expressive event mechanism could eliminate the need for polling entirely. In the case of network traffic, for example, one might request an event for every kilobyte of network activity.

All data are stored locally in a timestamped key-value store for periodic upload to a central server. By default Device Analyzer attempts to minimise the impact of this by scheduling uploads when the device is charging with a 802.11 connection available.

Timestamps Simple event timestamping using the current system time is inadequate in a number of cases. System time is subject to discontinuities, such as when the user manually changes the time on their phone or when the cell network broadcasts a correction. Furthermore, local time is not always available: we see that for some seconds after a system upgrade the returned dates are in the 1980s! To provide dependable timestamps in the face of such disruptions, we use the system uptime as our reference frame as it is guaranteed to not jump backwards or forwards. Changes to the mapping of uptime to local time are logged in the event stream.

Viewing Data Data can be viewed on the phone itself and on the web after it was uploaded. At the moment the data displayed on the phone is designed to show the user which kinds of data are collected and does not contain a history. This will be addressed in future versions. On the website users can compare their historical records with those of the entire population and get more detailed information, including graphs of network traffic and other variables.

Hashing and Aliases Rather than uploading identifying user data such as phone numbers or 802.11 network addresses in plain text, they are hashed with a device-specific salt to maintain user privacy. Users can choose to assign private nicknames to these anonymous values so that they can identify them when they log in to the website. The dataset itself contains only the hashed values.

Privacy Concerns We plan to periodically release the collected anonymous dataset where participants have given us permission to do so. This raises privacy concerns that need to be addressed before such a comprehensive and potentially invasive dataset can be released. As a first measure we will only publish data that was collected at least three months ago, giving participants time to view the data they contributed. If they are concerned about having their data released they can opt-out of any further data collection and erase all unpublished data from our server. Furthermore, while physical location of participants could be an asset of our dataset, such data would need to be handled with great care, as disclosure of even approximate home/work location pairs could lead to deanonymization of participants [2].

Conclusion A rich dataset of smart phone usage for a large population has undeniable uses for device owners as well as providers and manufacturers. However, collecting such detailed data raises privacy concerns even if the data are never uploaded if an adversary gains physical access to the device. We are working to address privacy issues which arise sufficiently for the scope of a research dataset where informed consent can be assumed. However, when considering a general deployment by a manufacturer, further protection may be required. It is an open question at this time how such a protection could look like.

We plan to release Device Analyzer on the Android Market this year and are currently conducting a beta test. Please do contact us if you wish to participate.

Acknowledgements

This work was supported by the University of Cambridge Computer Laboratory Premium Studentship scheme, a Google focussed research award and the EPSRC Standard Research Grant EP/P505445/1.

References

[1] N. Eagle and A. S. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, Nov. 2005.

[2] P. Golle and K. Partridge. On the Anonymity of Home/Work Location Pairs. *Proceedings of the 7th International Conference on Pervasive Computing*, 5538:390–397, 2009.