



The 11th International Conference on Mobile Systems and Pervasive Computing
(MobiSPC-2014)

A handset-centric view of smartphone application use

Juwel Rana^{a,*}, Johannes Bjelland^a, Thomas Couronne^a, Pål Sundsøy^a, Daniel Wagner^b,
Andrew Rice^b

^aTelenor Research, Snarøyveien 30, N-1331,
Fornebu, Norway

^bUniversity of Cambridge, Computer Laboratory,
15 JJ Thomson Avenue, Cambridge, UK

Abstract

Studying the use of applications on smart phones is important for developers, handset designers and network operators. We conducted a study on Android devices by installing an instrumentation application, Device Analyzer, on participants' handsets. Over a 4 month period we collected 10.9 billion records from 674 different users. In this paper we describe how to use the *research study* features of Device Analyzer to control participant selection and to access information (with consent) that is withheld for privacy reasons from the main dataset. We describe our data processing architecture and the steps required to preformat and analyse the data. Our data contains 3329 distinct applications (from the Google Play store) but despite this, on average, a user makes use of only 8 unique applications in a week. Almost 100% of our users make use of some email application on their phone. Fewer users (85%) made use of the Facebook application but 4–5 times more frequently than for email with sessions lasting almost twice as long. We also investigated whether different applications have correlated usage using a network analysis and a principal component analysis. We see that application usage tends to correlate by vendor more than by activity. This is potentially due to vendors integrating or cross-promoting services between applications.

© 2014 Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

Selection and peer-review under responsibility of Conference Program Chairs

Keywords: Android, smartphone, mobile apps, app usage analytics, big data;

1. Introduction

Mobile applications play a central role in the usage of modern smartphones. According to Business Insider 22% of the global population will own smartphones by the end of 2013². With the growing number of smartphone users in Europe and Asia, the market for app developers is getting bigger and apps are added and updated on a daily basis.

Understanding the usage of mobile apps is important for developers attempting to build better applications, for device manufacturers designing new handsets and for network operators attempting to provide a competitive high-quality service to subscribers.

* Corresponding author. Tel.: +47-917-966-21.
E-mail address: juwel.rana@telenor.com

The most direct method to capture application usage is to integrate analytics libraries (such as Google Analytics for Mobile) into applications themselves. This provides detailed information to the developer about their own application but cannot provide a broader view of the use of the handset as a whole. Other researchers have worked from a network perspective using Call Data Records (CDRs)^{9,8,10}. This provides behavioral information about huge numbers of users but only captures interaction with the network itself—which offers a limited view of handset use as smartphones become more and more general computing devices. The final approach is to install dedicated data collection software on a handset, thus getting a detailed view of usage but with fewer users than accessible through CDR-based studies. By installing a monitoring app that runs in the background, data is collected unobtrusively from the smartphone OS without any user input needed. In addition to probing user interactions like starting and stopping apps, making calls and sending SMS, the monitoring applications can also monitor a whole range of sensors such as GPS, accelerometers, gyroscopes, magnetometers and proximity sensors. Researchers have built a variety of datasets and tools to investigate particular questions in this way^{5,3}. In this paper we follow this approach making use of the Device Analyzer application¹. Device Analyzer is somewhat unique in that the project has been explicitly designed to make it possible to share collected data with other researchers and to support adding external data sources for researchers who run studies using the tool. To this date the Device Analyzer Project dataset contains 100 billion data points recording handset usage from over 18,000 individuals globally and a subset of this dataset is available for free to other researchers to use. Details are available on the project website¹. However, in this study we want study app-usage locally in Norway and also collect demographics for the participants.

We use Device Analyzer to investigate application usage on Android handsets and we collected 10.9 billion data points from 674 users. The purpose of this paper is to share our preliminary findings and to inform other researchers considering this kind of study. We describe the research study mechanism in Device Analyzer and the steps we needed to take to normalise and analyse the data collected. Simple analysis of application usage shows that almost 100% of users make use of an email application on their phone. Fewer users (85%) made use of the Facebook application but 4–5 times more frequently than for email with sessions lasting almost twice as long. We also investigated whether different applications have correlated usage and show the results of a network analysis and principal component analysis for this. We see that application usage tends to correlate by vendor more than by activity. This is potentially due to vendors integrating or cross-promoting services between applications.

1.1. Data collection

We collected device log data from 674 Android-based smartphone users in Norway. The data collection period was from September 2013 to January 2014. The trial users were recruited through online social media sources by posting a link containing information about the study, and also the procedure for installing Device Analyzer. As an incentive, participants were rewarded with chances to win in a lottery. In the sample group, we found 75% male and 25% female with a median age 36.

The app collects a rich set of features, e.g. process logs, location, hardware settings, power consumption, connectivity and signal strengths. A detailed description of the collected data is available on the project website.

Privacy is a major concern with studies such as this. Device Analyzer de-identifies user data at the point of collection and also provides safe defaults concerning the sharing of potentially sensitive data. The full device analyzer dataset (18,000 users to date) is available to researchers; the default privacy settings however mean that application names and location information are not included. Additionally, there is no other information (such as demographics) available about the contributors to the dataset.

In order to help researchers who require this extra information Device Analyzer provides a research study feature. We designed a study and performed own research ethics review before recruiting our set of participants. These participants then used the normal Device Analyzer application and entered a unique code which we delivered through SMS. This code serves as a pseudonym and means that the user's data is automatically collected and shared with the study organisers. To maintain user privacy, the links between the phone numbers and study IDs were erased after the study period.

¹ <http://deviceanalyzer.cl.cam.ac.uk>

Fig. 1. Summary of collected data

Facts	Values	Unit
Number of participants	674	Users
Total rows parsed	10.9	Billion
Number of rows generated	16	Million/user
Average number of unique apps	8	per User/Week
App screen time (Median)	88	Minutes per User/Day
Number of apps started	38	per User/Day

The sign-up mechanism for our study made use of an online survey to collect demographic information which would otherwise not be available through Device Analyzer. We find that our sample is skewed towards male users (75% of participants). The participants are distributed all over Norway. We also asked participants for their consent to include their Telecom Billing Data. Our pilot group was surprisingly willing to share this, as 98% of the participants answered yes to this question. We plan to make use of this information in future studies comparing the network-centric view of a user to the device-centric one.

The collected data was around a terabyte in size and consisted of 10.9 billion rows. A full list of the information collected by Device Analyzer is available from the website.² Figure 1 shows a summary of the data we collected. Interestingly we see that most users stick to a relatively small number of core applications with an average of 8 unique applications used per week but spend an average of 177 minutes per day with applications in the foreground on their device. In comparison, Oliver found in his 2010 study that BlackBerry users spent an average of over 100 minutes per day on their devices¹¹. Observing median, Oliver reports a median time of 78.6 minutes, while we found 88 minutes. We observe that the distribution is quite skewed with some extreme users.

The rest of the paper is structured as follows, Section 2 provides data preparation methods, Section 3 describes high level architecture for massive smartphone log analysis and then, Section 4 provides analysis results. Finally, Section 5 concludes the paper.

2. Data Preparation

The raw data stream is complex and needed to be preprocessed before the analysis step. There are many different parameters in the data and most of these correlate with each other. Therefore it is important as well as complicated to provide a structure of this.

2.1. App Usage Dataset Preparation

Android application process names appear as package names rather than the human-readable name of the application in the raw data. For example, we found Facebook app uses `com.facebook.katana` as URI and Facebook messenger uses `com.facebook.orca` as URI. We converted package names back to readable ones through the use of a Python script which performs a lookup against the Google Play online app store³. Out of the 7722 distinct applications used we found only 3329 of them available in Google Play. The vast majority of the remainder were vendor specific applications pre-installed on the handset.

In order to reduce battery usage Device Analyzer only polls the lists of running processes with a fixed 5 minute interval. In addition to being coarse-grained this provides little information about the state of the application (background, foreground, service). The ‘High Frequency’ sampling measure is provided to address this: in this mode the ‘last started app’ is stored with a frequency of 2 Hz. This mode is randomly enabled for 10% of screen-on periods. We extrapolate this value to count application starts and estimate usage times. We leave the task of determining the accuracy of this approach to a later study.

² <http://deviceanalyzer.cl.cam.ac.uk/keyValuePair.htm>

³ <https://play.google.com/store>

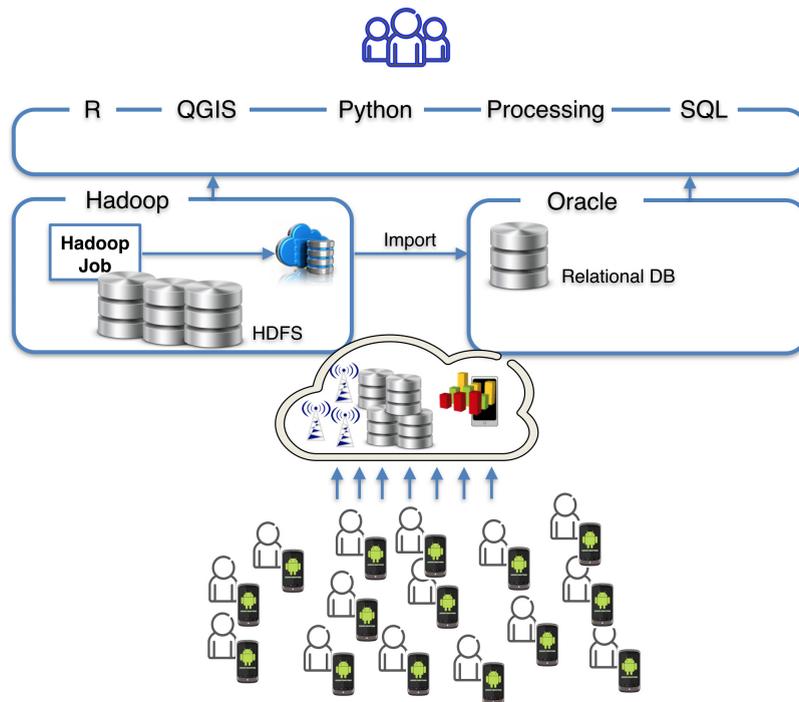


Fig. 2. High level view of MSLA Architecture

2.2. Correlating Location with App Usage Dataset

Device Analyzer collects coarse-grained location (latitude and longitude) of the device every 5 minutes (with user consent). This enables us to group sessions of app usage with the temporally closest latitude and longitude of the device, allowing us to approximate the location of app usage throughout our study.

3. High Level Architecture for Massive Smart Log Analytics

The size of our data logs and the variety of information contained within them is a big data problem⁴. We built a generalised architecture called "Massive Smartphone Logs Analytics" - architecture (MSLA) (Figure 2), to perform the analysis of this data. The MSLA architecture governs the complete analysis process. Firstly, it collects data from Device Analyzer's cloud storage¹. At this stage, the data is semi-structured key-value paired which then must be preprocessed before any analysis can be performed. This batch processing is performed using Hadoop⁴. The Apache Hadoop software library allows massive smartphone data to process across clusters of computing nodes. In Hadoop, the batch file is stored in hadoop distributed file systems. After that, we performed map-reduce jobs on HDFS to prepare structured datasets for specific data analysis purposes. We import the outputs into an Oracle relational data storage for convenient querying. The Data in the Oracle data storage, is queried using SQL developer and visualized with tools such as QGIS⁵ or Processing⁶. The data is also analysed using the R statistical computing tools. For example, the visualization in Figure 4 is done using QGIS. All the data movement between the tools are handled using Linux and Python scripts.

⁴ <http://hadoop.apache.org/>

⁵ <http://www.qgis.org/>

⁶ <https://www.processing.org/>

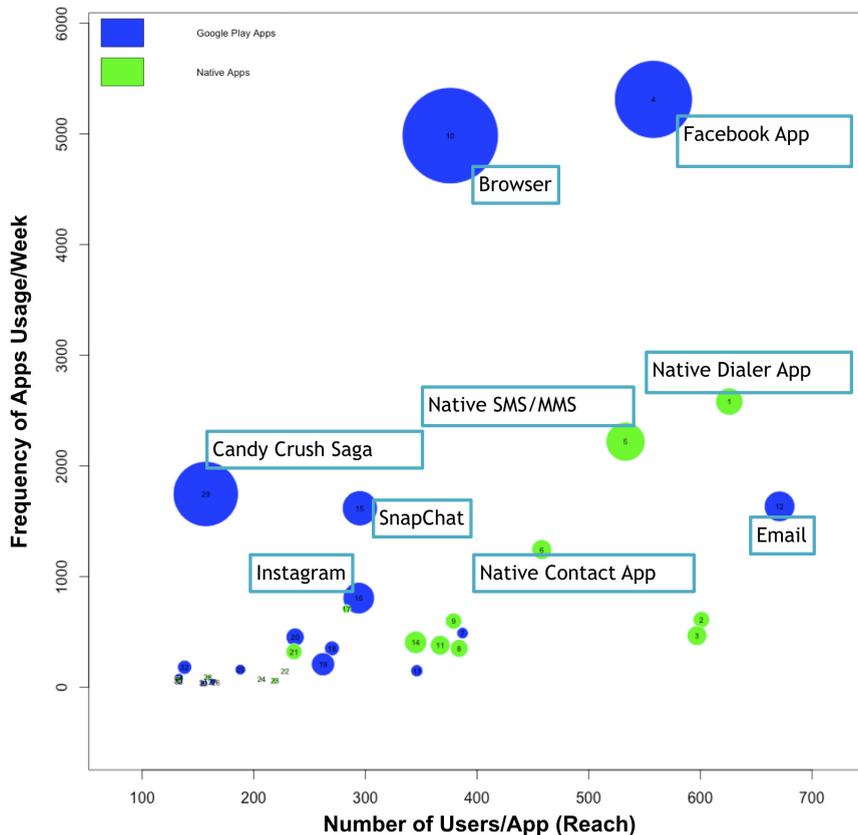


Fig. 3. Application usage within our dataset

4. Analysis Results

4.1. Application Popularity

We classified top apps in terms of number of initiations, duration of usage, and number of users. Figure 3 summarises our results. The X-axis corresponds to number of users per application whilst the Y-axis corresponds to frequency of usage per week. The bubble size corresponds to the average session length. It is interesting to see that Facebook is the most frequently used app, while web browsers are used for longer intervals at a time. We also note that during our study many users never opened a web browser. Those users who interacted with Facebook or a web browser opened the applications almost twice as often as apps that correspond with more traditional (non-smartphone) use, such as the Dialer or text messaging. Almost every user used their device for emailing but interactions were comparatively brief.

We identified thirty categories for 3329 applications from the Google Play store. Pre-installed apps provided by the device manufacturer or apps downloaded outside from Google Play were omitted. Despite the diminishing use of traditional phone activities as outlined above we see that smart phones are still principally used for communication. Social and Communication apps dominate in both frequency and duration, covering 60% of the total app usage. Apart from Social and Communication, other popular categories are Tools, Productivity, Music, Finance, Casual, News, Games-Puzzle, and Entertainment.

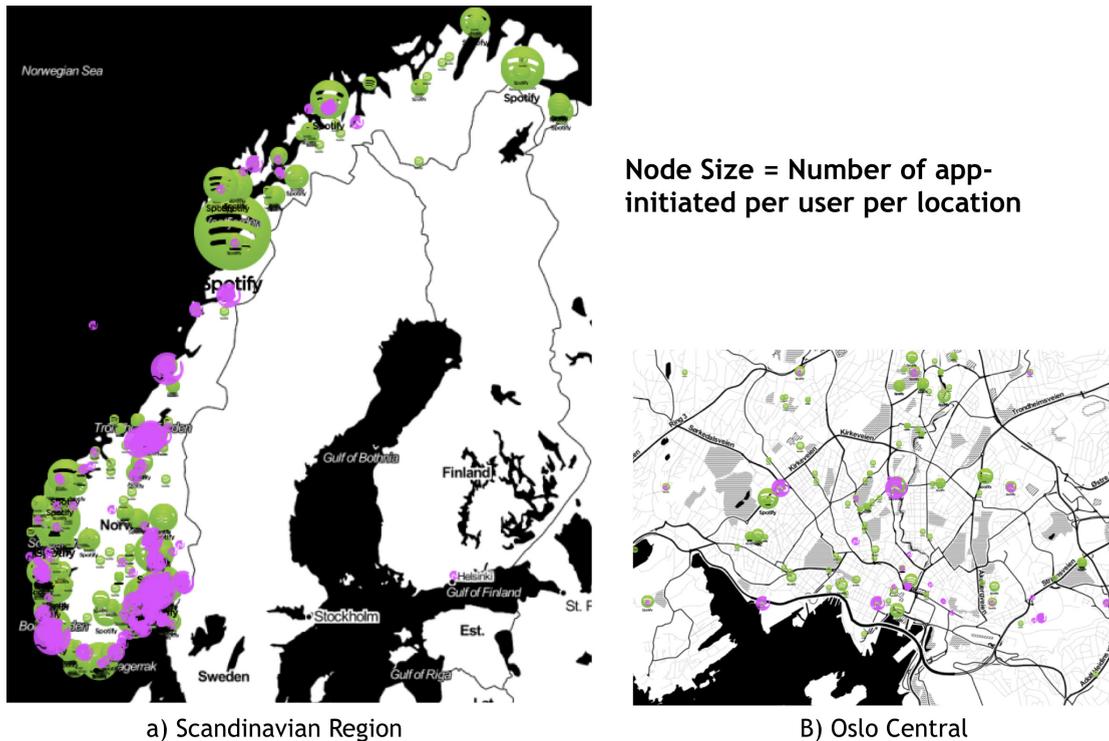


Fig. 4. Where are WIMP & Spotify Started? Scaling of node size represents the average number of app-initializations per location per user

4.2. Location of Apps Initiation

Mobile location data is valuable for providing personalized services to the mobile device users⁶. Similarly, the location of application usage can show interesting usage patterns⁷. Based on our experimental dataset, we visualize the locations where apps were started. The visualization results help in decision making; for instance, we can extract information about frequently used mobile apps while using public transportation, which may be of use when rolling out WiFi hot-spots on public transportation vehicles. Similarly, visualizing where the various categories of apps are used may highlight mobile information needs in particular locations¹².

Figure 4 shows the location of Wimp and Spotify usage during the experimental time period. Spotify users are distributed more evenly across the entirety of Norway, while Wimp users are mostly active in the southern part of Norway. A similar study could be run to gauge the geographical distribution of other forms of apps—like Google Maps or Facebook. The process may help data providers when allocating resources to particular regions for better network coverage. Streaming apps like Spotify, Wimp and Youtube typically require a higher bandwidth than other apps. This type of highly granular data showcases some of the advantages of on-device data collection and allows telecom operators to analyse data that would typically not be available with traditional data collection methods.

4.3. Network-view of application use

Figure 5 shows a network visualizing weighted co-usage of apps. The vertices of the graph represent apps; edges indicate that two apps are used by the same user in the same time window (4 hours).

We find that Eigenvector Centrality (EVC) correlates well with application popularity, and two applications located nearby have higher probability to be used in same context. Despite the visual noise in the network we see that some apps are clustered closer together. By analyzing the graph we find that the top 5 apps—in terms of degree (# of co-use connections) and EVC—are Facebook, Google Chrome, Snapchat, Instagram and Gmail. This means that they are

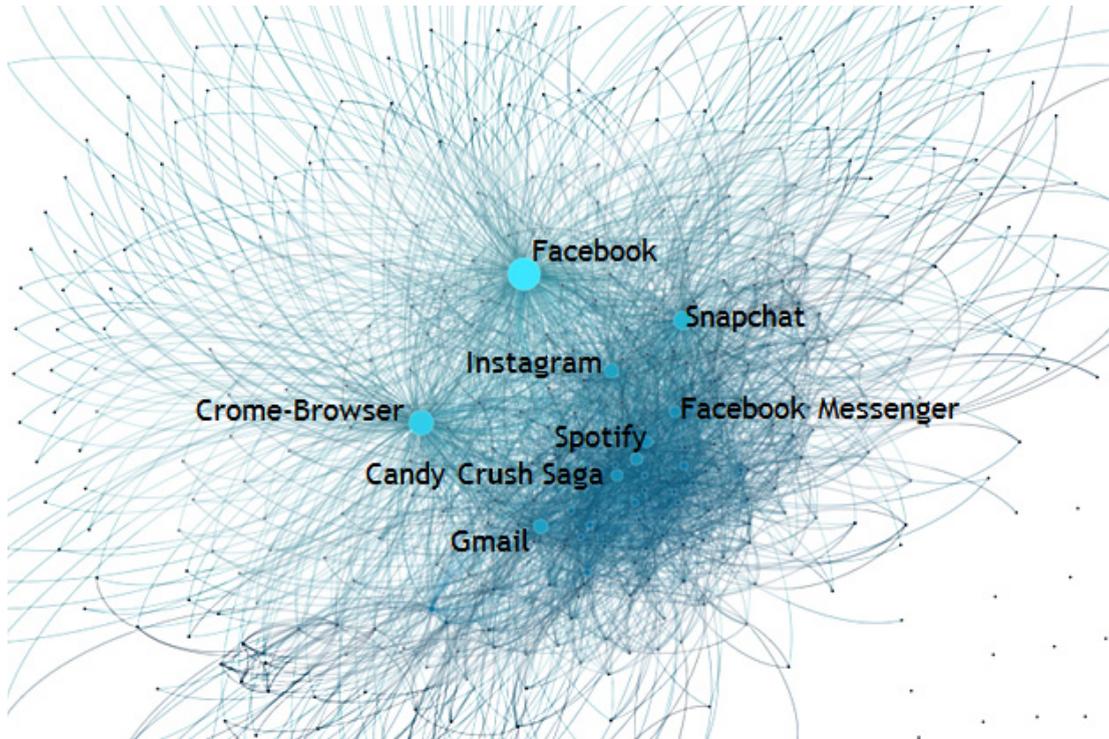


Fig. 5. Mobile application co-use network

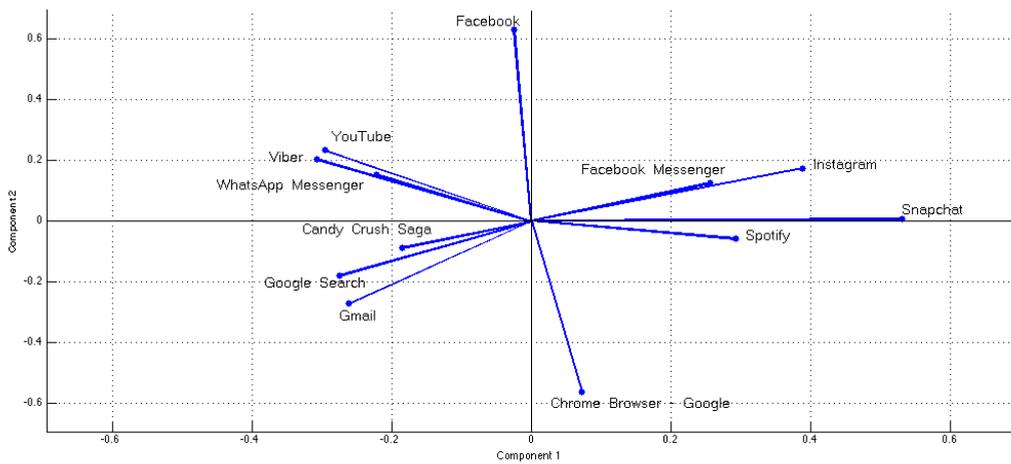


Fig. 6. A Principal Component Analysis of top applications showing correlated use

not only popular in isolation, but also used heavily together with other apps. They are also the most popular in terms of duration, frequency and number of users.

Finally, we performed a principal component analysis (PCA) on the dataset. Figure 6 shows the result of PCA considering the top most used applications. To do so, we select the correlation circle to understand how variables are interdependent. Our first observation is that in the south-west quadrant we find mostly Google products while in north-east quadrant we find dominantly Facebook products (Messenger, Instagram, Snapchat.) These grouped

products represent different, but related, activities. One possible explanation for this grouping of applications is that the different services are cross-promoted by providers.

We also observe that the use of Viber and Whatsapp are strongly linked, and seem to be independent to the use of the previously mentioned Google versus Facebook products. Last but not least, Facebook and Chrome Browser are negatively-correlated which means Facebook application users rarely use Chrome Browser and vice versa. The Spotify app is associated with Snapchat and Instagram use but it is negatively-correlated with the use of the Youtube app (Google's product). It will be interesting to see what effect Facebook's acquisition of Whatsapp has on this breakdown in future.

5. Conclusions

Developers, device builders and network providers could all benefit from a better understanding of application usage. We used the Device Analyzer tool to run our own research study to collect detailed information of application used from 674 subscribers. In this paper we described how this process works and how we augmented the collected data using a questionnaire and pseudonymous identifiers. We gave an overview of the resulting dataset and discussed the steps we put in place to filter through and analyze the considerable amount of data in these logs.

Our main goal of this paper is to explore various methods for analyzing and understanding massive smartphone logs. We have shown some of the potential of the data such as associating location with app usage. This allows us to get a better understanding of the context of app usage - e.g. regional differences and travel usage like commuting versus app usage at home. We reserve a detailed analysis of contextual app usage for future work.

The vast majority of our participants consented to the analysis of their Telecom Billing Data in tandem with their handset usage. This analysis will form the basis of our future work. A deeper insight into application networks might be possible using community detection algorithms which can simplify the visually noisy picture shown in Figure 5. We also wish to analyze in more depth the spatial-temporal dimensions of application use and how they change over time.

6. Acknowledgments

We want to thank Geoffrey Canright and Taimur Qureshi for bringing insights and good discussions through the whole project. Special thank goes to Telenor Norway for helping in recruiting pilot participants.

References

1. Wagner D., Rice A., Beresford A. Device Analyzer: Understanding smartphone usage *10th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*; 2013.
2. <http://www.businessinsider.com.au/smartphone-and-tablet-penetration-2013-10/>
3. Yu P., Au Yeung C. M. App mining: finding the real value of mobile applications. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*; 2014.
4. Musolesi M. Big Mobile Data Mining: Good or Evil?. *Internet Computing, IEEE*; 2014.
5. Qian F., Wang Z., Gerber A., Mao Z., Sen S., Spatscheck O. Profiling resource usage for mobile applications: a cross-layer approach. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*; 2011.
6. Qu Y., Zhang J. Trade area analysis using user generated mobile location data. In *Proceedings of the 22nd international conference on World Wide Web*; 2013.
7. Bhmer M., Hecht B., Schning J., Krger A., Bauer G. 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*; 2011.
8. Sundsøy P., Bjelland J., Iqbal A., Pentland A., Montjoye A. Big Data-Driven Marketing: How machine learning outperforms marketers gut-feeling. *LNCIS Volume 8393*, pp 367-374; 2014.
9. Sundsøy P., Bjelland J., Engø-Monsen K., Canright G., Ling R. Comparing and visualizing the social spreading of products on a large-scale social network. *The influence on Technology on Social Network Analysis and Mining*. Tanel Ozzyer et.al, XXIII, 643 p. 262 illus, Springer; 2013.
10. Bjelland, J., Canright, G., Engø-Monsen, K., Sundsøy, P., Ling R. A Social Network Study of the Apple vs. Android Smartphone Battle *Advances in Social Networks Analysis and Mining (ASONAM)*, vol., no., pp.983,987, 26-29; 2012.
11. Oliver E., The challenges in large-scale smartphone user studies. In *Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement - HotPlanet*; 2010.
12. Church K., Smyth B. Understanding mobile information needs. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services - MobileHCI 08*; 2008.