Towards rich mobile phone datasets: Lausanne data collection campaign

Niko Kiukkonen[†], Jan Blom[*], Olivier Dousse[*], Daniel Gatica-Perez[‡] and Juha Laurila[*]
[*]Nokia Research Center, PSE-C, 1015 Lausanne EPFL, Switzerland
Email: *first.last*@nokia.com
[†]Nokia Research Center, Itämerenkatu 11-13, 00180 Helsinki, Finland
Email: niko.kiukkonen@nokia.com
[‡]IDIAP Research Institute, Rue Marconi 19, CP 592, 1920 Martigny, Switzerland
Email: gatica@idiap.ch

***Abstract*—Mobile phones have recently been used to collect large-scale continuous data about human behavior. In a paradigm known as people centric sensing, users are not only the carriers of sensing devices, but also the sources and consumers of sensed events. This paper describes a data collection campaign wherein Nokia N95 phones are allocated to a heterogeneous sample of nearly 170 participants from Lausanne, a mid-tier city in Switzerland, to be used over a period of one year. The data collection software runs on the background of the phones in a non-intrusive manner, yielding data on modalities such as social interaction and spatial behavior. The main motivations for organizing a new campaign on top of the ones that have been successfully conducted in the past are the following: First, in comparison to the Reality Mining data, generated in 2004-2005, the present data set is expected to provide a richer means to study location attributes, in particular, because today's mobile phones are more powerful and equipped with more sensors. Second, we aim to recruit a heterogeneous set of participants, comprising family and leisure related social networks in addition to organizationally driven ones. This paper provides a methodological description of the project and shows the potential of the resulting data set in terms of illuminating multiple aspects of human behavior.**

I. Introduction

In the recent years, the penetration of mobile phones has reached particularly high levels. Besides, the number of sensors embedded in these devices has also dramatically increased. The combination of these two trends has enabled a new kind of research called *people centric sensing* [1]. In this approach, sensors in mobile phones and other wireless devices are used to collect large quantities of continuous measurements about their users. The data collected in this way captures various types of behaviors, ranging from intra-personal to inter-personal level. Computational techniques can be applied on the resulting data sets, enabling high level attributes to emerge from the data.

People centric sensing can lead to a qualitatively new type of approach to studying social systems, such as social networks and organizations. Lazer et al. [2] refer to the capacity to collect and analyze vast amounts of socially grounded data as computational social science, drawing an analogy to the rise of cognitive sciences a few decades ago. Raento et al. [3] claim high ecological validity for harnessing data collection capabilities of smart phones, due to the mobile phone being accessible to large parts of the general population and to the non-intrusive nature of data collection applications running quietly on the background of the mobile phone.

One of the most widely cited people centric sensing activities is the Reality Mining project [4]. Since the collection of the data set in 2002, it has become a popular resource for the people centric sensing community, inspiring research related to, e.g., social network patterns [5], behavioral regularities [6], and behavioral entropy [7]. While the scientific contributions enabled by the data set have undoubtedly been prevalent, wireless technology has progressed since the campaign. New types of sensors appear in todays handsets, adding to the richness of the modalities that can be collected by the experimental phones. The density of wireless networks, especially in urban areas, has also increased due to the proliferation of Bluetooth enabled devices and WLAN access points; the sensing of stationary as well as moving wireless devices yields more data now than it did half a decade ago.

Hence in order to push the possibilities of data analysis further, generating a new data set using people centric sensing methods has become justified. The key dimension along which progress has been made since the early 2000s is location. In Reality Mining, inferences about the position and movement of the participants were limited to using cell ID information available to the mobile phones of the participants. However, for a variety of reasons, the geo locations of cell tower ID's are not always publically available. Also Bluetooth beacons detected by the experimental phones were utilized to improve location accuracy. The smart phones of today provide two new sensors, namely GPS and accelerometer, increasing the precision of outdoor based positioning as well as mining of the movement patterns of an individual. The ability to analyze spatial data collected in indoor environments has also improved because of the increasing number of WLAN access points found in buildings and the phones' capability to detect them.

This paper describes a people centric sensing campaign designed to provide a comprehensive view of the spatial-social-temporal environment of the participating individuals. The increased resolution was thought to be particularly linked to yielding richer location based data, in comparison to the Reality Mining project, which operated on a more narrow range of attributes. Another motivation for deploying a new campaign was to be able to study the behavior of a more heteregeneous population. Reality Mining was based on data collected by individuals affiliated with MIT. The occupational range was limited to staff, faculty members as well as students. Our target was to introduce more variety to the set of individuals whose behavior is captured. The assumption was that by recruiting beyond an academic population, the number and nature of patterns that can be extracted should get more diversified.

In addition to the above data driven and sampling related justifications, there was also a pragmatic reason for setting up a new people centric sensing campaign. Previously cited research in this field has been conducted by academic parties [8], [9]; a corporate research lab from the mobile industry is behind the present campaign.

A. Corporate perspective to people centric sensing

People centric sensing can contribute to behavioral theories concerning individuals and social systems. However, this paradigm is also important from an industrial point of view. The increased sensing capabilities of mobile devices, as well as the possibility of uploading data to a server in real time fashion, are utilized in context aware services. Google Latitude [10] and Nokia Maps [11], for instance, offer commercial services showing the location as well as status of members of a users' social network on the map, in a real time fashion. Citysense [12] is a mobile application that operates on an aggregate level, showing the overall activity level of the city, top activity hot spots, and places with unexpectedly high activity. It can thus be asserted that people are now, and at once, the carriers of sensing devices, and the sources and consumers of sensed events [8].

Both theoretical and pragmatically driven research questions can be applied on the present campaign. Consequently, three distinct levels of analysis concerning the resulting data set emerge: (1) behavioral, (2) consumer centric; (3) user centric. Each of these levels views the campaign members from a different angle.

First, the behavioral level aims to discover novel aspects of human behavior using the computational social science [2] paradigm. The key value lies in the ability to capture continuous data about behavior in non-intrusive manner, consequently enabling powerful statistical techniques to be used for data analysis. This level is mainly theory driven; it can be considered as basic research contributing to the general knowledge of the scientific community.

Second, the campaign members can also be considered as mobile phone users, rendering the perception of the sample to that of consumers. Collecting data about how mobile phones are used in the population can reveal important insights about the interrelation between the consumer, the product and the context of use. This is valuable at the commercial level, allowing, e.g., marketing and product development related inferences to be made, subsequently increasing the fit of the product offering to the needs of the consumer base.

Third, in the user centric level, campaign members are engaged in reflective activities converging on the notion of people centric sensing, and the potential value thereof. User centric design methods [13] are applied and the participants are viewed as co-creators. The campaign familiarizes the participants with the nature of people sensing, enabling articulation of perceptions and desires concerning context-aware services in an ecologically valid way.

While most of the prior people centric sensing literature adopts a rather theoretical level of analysis, the entire theory-practice continuum is relevant for a corporate research entity such as Nokia Research Center. The stakeholders of the output of a corporate research lab include not only academia but also corporate entities.

In addition to the industry background influencing the selection of research angles, it also led to the privacy policy incorporated by Nokia in provisioning of digital services being adopted for the data collection campaign. The privacy related dynamics will be elaborated on in the next section.

In line with the three motivating factors discussed above, the contributions of this paper are as follows. We aim to provide a comprehensive description of the data collection campaign, ranging from the method used to highlighting the potential of the resulting data set. We also aim to underline the corporate twist inherent to the design of the campaign as well as to analysis of the data, as it distinguishes the project from prior people centric sensing research.

The next section describes the method of the project from three perspectives: participants, technology and privacy. We then describe the scope and magnitude of data collected so far, to highlight the potential of the data set.

II. Data collection method

The data is collected using smart phones equipped with special software aiming to make the data collection invisible to the participants while optimizing the ratio between data collected and power consumed. The non-intrusiveness was seen critical in order to make it convenient for participants to stay within the campaign for the target period of twelve months. The data collected is stored to the device and automatically uploaded to the server handling the data when the device detects a known WLAN access point. The server receiving the data twice a day from all participants post-processes and builds a SQL database from the incoming data. The database can then be accessed to perform analysis on the (anonymized) data and to visualize it to the participants. The schematic view of the full data collection system is presented in Fig. 1. The different data modalities collected with the client can be categorized as follows:

a) Social interaction data: Social interaction is inferred from call logs, short message logs and Bluetooth scanning results. In addition, we can use information from acoustic environment samples to detect the devices sharing the same acoustic space at any given time. Together these parameters can reveal the events where the persons in question are

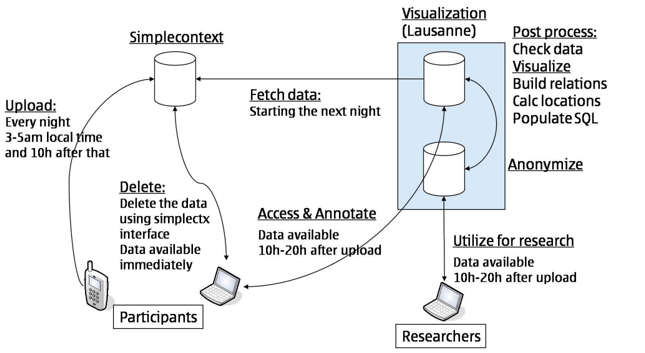

Fig. 1. Data flow between the experimental mobile phones and the databases

interacting with each other. This information can be annotated further with the location information to understand the motivations behind the proximity.

b) Location data: Location is determined based on GPS (when available), cellular network information, and WLAN access point information (when available).

c) Media creation and usage data: Information is captured concerning locations where images have been captured, video shot or music played.

d) Behavioral data: Information is received concerning application usage, activity detection based on acceleration sensor, and regular device usage statistics based on call and short message logs. The locations visited and transportation means can be derived from location data. This data can be complemented with the help of questionnaires administered for the participants.

A. Data collection client software

Nokia N95 handsets are used to collect the data and the process is designed to be non-intrusive. The data collection software was designed to meet the operating times of one full day enabling participants to use their device normally during the day and charge the batteries over the night. This was a challenging goal given that, simultaneously, the target is to collect rich and comprehensive data. We identified the following three strategies in regard to power consumption: a) Limit the sampling to be infrequent but utilize the long duration of the campaign to collect enough data to fill the gaps; b) Use high sampling frequency until the battery level reaches a predetermined threshold and stop sampling after that; c) Have the data collection client software analyze the conditions and tune the sampling accordingly. We selected option c) as the basis for client software development. This approach optimizes the power consumption and the amount of collected data throughout the operating time of the device.

The data collection client is based on a state machine running through 13 possible states. The range of parameters sampled as well as the sampling frequency varies between the states. There is a potential of an error in misinterpreting the prevailing conditions and consequently selecting an inappropriate state for the sampling. To minimize this risk our client uses multiple inputs in each state to drive the state transitions. This way, error in any individual contextual parameter will not trigger the state transition. The states are changed based on motion information from acceleration sensor, radio scanning results consisting of WLAN, cellular network, and GPS information, and the device state information (for instance, whether or not the battery is connected to a charger).

State	Power consumption	Operating time
All on	1.4W	2.8h
outdoor	0.54W	8.5h
known WLAN	0.2W	23h

TABLE I
POWER CONSUMPTION AND MAXIMUM OPERATING TIMES PRODUCED IN SELECTED OPERATING MODES.

The effect of the state machine based operating mode to the battery performance is dramatic. In comparison to a case where the sampling is continuous (All on) the operating times can be improved significantly both when being outdoors and when being at home or in the office (under known WLAN). The same events can still be reliably collected (Table I). The most relevant states used, and the modules the client records in these states are listed in Table II in the order of occurrence.

The sampling of the client and the state transition logic is described below for the most frequently occurring states.

- In the *Known WLAN* state the client software operates under WLAN access point it has detected recently, for cumulatively counting long enough a time. The *Known WLAN* state is implemented to save power based on the knowledge of the location of the participant tracked with the WLAN access points. During the implementation and testing we identified that the scanning of the WLAN is not very reliable causing the device to miss or lose randomly some or all visible WLAN access points. To fight against this limitation in the WLAN implementation, the client records the acceleration values in addition to WLAN scanning to detect possible movement and thus improve the reliability. If the known WLAN access point is lost and acceleration values are above movement threshold, the client changes its state. The client tags automatically the known WLAN access points with GPS coordinates if the location information is available roughly simultaneously with the WLAN scanning results. This facilitates subsequent location related processing on the server side.
- *Indoor Mobile* is a transition state where the client software detects whether the user is transitioning to an outdoor space or just changing the location indoors. The software keeps on tracking the WLAN networks to detect previously encountered ones when they re-appear. Also previously unencountered access points are recorded. The GPS tracking is turned on after a predefined period to find out if the participant has moved outdoors. The acceleration is monitored to trigger the state transition to *Stationary* state if the movement ends.
- In the *Outdoor mobile* state the client accesses GPS position information on a continuous basis. Additionally

State	Modules sampled and sampling interval	Approx. occurrence
Known WLAN	WLAN (120s), BT (60s), Cell (60s), Accel. (60s)*	40%
Stationary	WLAN (600s), BT (180s), Cell (60s), Accel.(60s)*	15%
Outdoor mobile	WLAN (60* or 300s), BT (180s), Cell(60s), GPS(cont.), Accel.(300s)	10%
Plugged in	All (60s)	10%
Known WLAN lost	WLAN (60s), BT (60s), Cell (60s)	8%
Battery low	BT (300s), WLAN (900s), Cell (60s)	8%
Indoor mobile	BT (300s), WLAN (60s), Cell (60s), Accel. (60s)*	1%
Other states	Misc.	8%

TABLE II
THE MOST RELEVANT STATES, WHAT MODULES ARE SAMPLED, AND THE APPROXIMATE FREQUENCY OF THE STATE IN THE CURRENT DATA. IN ALL STATES CALL & SMS LOGS, APPS & MEDIA USE, ACOUSTIC ENVIRONMENT INFO ARE SAMPLED.*) IN THESE STATES THE ACCELERATION IS ONLY USED TO TRIGGER THE STATE TRANSITION AND IT IS NOT STORED

the WLAN access points are scanned relative to the detected speed. When the speed is high the scanning is not performed at all. The cellular network parameters are recorded to enable position tracking based on cellular data in cases where the GPS signal is lost. Bluetooth devices are scanned to detect the devices (and persons) in proximity and traveling together. Acceleration values can be used to detect modalities, such as vehicle types.

- In the *Stationary* state, GPS location tracking is discontinued. However, the cellular and WLAN network parameters are regularly scanned to detect possible movements. Similarly acceleration sensor values are monitored. The sampling of WLAN networks can be done irregularly because we know the device is stationary and WLAN information should remain the same. Bluetooth devices are regularly scanned to detect new devices appearing in the proximity of the user.
- The *Battery low* state occurs when battery level reaches critically low values (15%). The sampling is therefore limited to save the remaining battery capacity.
- The *Plugged-in* state is used when the device is connected to charger. Direct access to energy enables the client to turn on all sensors and sample them frequently. Longer charging time is the downside of this state but it is rarely seen as a problem for participants because charging often happens at night.

As can be seen from Table II, the client (and hopefully participants as well) is statistically most of the time in the *known WLAN* state and pretty infrequently in states where GPS is activated, thereby optimizing the power consumption.

The call and SMS logs, as well as application and media information, are sampled in all states. The acoustic environment information is collected from the ones who have consented to this. The data is sampled using the build-in device microphone and sampled every 10 minutes for 30 seconds at a time and only during the daytime.

The data collection client samples the environment according to the state it currently is in. The sampled parameters may trigger a state transition. All samples are stored in the device and uploaded over known WLAN access points. The client also stores the current state helping the processing later on. To protect the privacy of the participants, the SMS logs and acoustic environment samples are reprocessed in the device before uploading. In addition, the client software automatically annotates the most frequently encountered WLAN access points with GPS coordinates and is later on capable of using these for positioning to further improve the power consumption.

B. Participants

A snowball sampling method was used for the data collection campaign, as the goal was to include a heterogeneous set of real life social networks in the campaign population. An initial set of individuals was chosen from Lausanne, a second-tier city in Switzerland, to serve as seeds for generating the sample. These start nodes were encouraged to recruit their friends, colleagues, and family members eventually leading to a population comprising real life social networks with individuals from mixed backgrounds. Due to its viral nature, recruitment was not a one-off event, but a continuous process spanning several months.

The enrollment sessions took place on an individual basis, in the premises of the research team running the campaign. The procedure for the enrollment session was as follows. The participant was asked to fill in consent forms as well as a detailed questionnaire assessing e.g. communication habits as well as the demographic background of the individual. The questionnaire also required the individual to name the members of one's social network that were also taking part in the study. This was hoped to facilitate establishing a ground truth in terms of who knows whom in the population. In the final part of the enrollment session, the experimental phone, Nokia N95, was handed out to the participant while also installing the personal SIM card of the participant on the handset. The web sites hosted at Nokia displaying the personal data of the individual, both in raw as well as visualized forms, were introduced to the participant.

The participants were required to consent to using the experimental phones as their primary phones in the course of the study. They were encouraged to stay in the campaign for a period of one year. Monetary reward for participation was not provided but data transfer costs caused by the data collection client were compensated at the fixed level of EUR 20 per month.

Despite the lack of explicit monetary reward, there were nevertheless other incentives to participate. Being able to upgrade to a high end mobile phone acted as a motivational factor for several individuals, who were using low cost or out of date phones up to then. The compensation provided for the data transfer costs acted as an incentive, too, since the amount given was enough to leave some free quota for using mobile Internet for personal purposes. The campaign was also promoted as a chance to contribute to the scientific goal of learning about human behavior. Hence for some participants, altruism acted as a motivating factor.

C. Privacy

Since an extensive amount of data is collected for each participating individual, privacy related aspects influenced the design of the campaign. On one hand, the research team has an ethical and legal obligation to ensure that the privacy of the individuals is respected. This requires a holistic approach, all the way from communicating the privacy policy to the participants, to taking appropriate measures at the back end to protect the anonymity of the participating individuals. On the other hand, since the participants are subject to continuous sensing, the research provides an ecologically valid way to study perception of participants towards privacy. Hence, it also becomes possible to conduct privacy related research in conjunction with the campaign. The two sides of privacy, ensuring privacy of participants and perceptions related to privacy, are discussed in further details here.

1) Ensuring privacy of participants: The guidelines and measures taken to protect the privacy of the participating individuals are strict, as they are in accordance with the general privacy policy of the corporation itself. This means e.g. getting consent from the participants to collect data from them for research purposes, informing participants about their rights with respect to the data, giving possibility to opt-out any time, and taking care of the data security. In addition to applying a strict procedure to protect the privacy of the participants, it was important to make the participants fully aware of the underlying privacy policy. This also contributed to the establishment of trust between the campaign participants and the research team running the campaign.

It was emphasized that the data is owned by the participants themselves and that each individual has the full right to decide what to do with their data. A dedicated web site allows accessing a personal profile wherein all the data records collected for any given individual can be viewed. The tool allows participants to delete some or all of their data. The campaign also offers an online visualization tool that helps the participants to view socio-temporal-spatial patterns inherent to their behavior. This tool can act as a kind of diary for the participants over the duration of the campaign and allows to add further information to the data as tags. An example of participant's view to his own data is presented in Figure 2.

In the design phase of the campaign, we made a clear decision not to collect any content information. For example, acoustic samples are collected as spectral coefficients which are additionally obfuscated to guarantee that the content of the discussions cannot be captured. Similarly, only statistical information is collected from SMS messages.

With our holistic privacy approach also technical means are needed to protect the privacy of the participants. One important element in this respect is the anonymization of the data. While the campaign participants can access their personal data in a raw, unprocessed format, the researchers will only be able to access the data that has gone through the automatic anonymization process. This includes replacing identifiable information with pseudonyms (names, phone numbers, identifiers of WLAN/Bluetooth nodes) and diluting the position accuracy at least around easily identifiable locations (like home or office). Further, in this kind of corporate-driven campaign the data security can be handled according to the company practices and standards. Therefore, no compromises are made in terms of risking the data flowing to the hands of non-authorized instances.

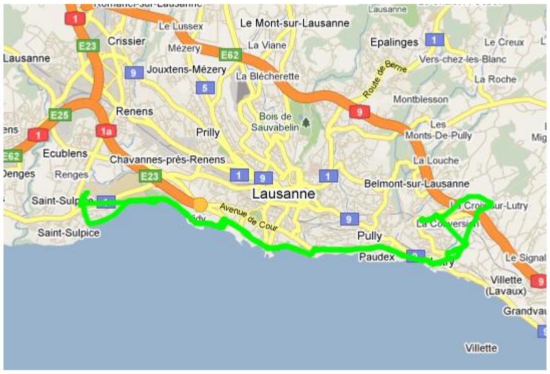

Fig. 2. Screenshot from the data collection Web site accessible to the campaign members. In this example visualization, the green pattern indicates the movement of one of the authors in the course of one day (the far ends represent home and the office).

The anonymization of this kind of research data always implies a trade-off between perfect privacy protection and value of the data to address different kind of research questions. That is, the anonymization should not go so far that the research value of the data gets too diluted. Also the accuracy requirement of the data might vary depending on the nature of the research question. For example, research questions related to transportation studies might require higher position accuracy than questions related to studying social systems. Therefore, final protection of the participants' privacy is arranged via non-technical means. Researchers will be only granted the right to access the anonymized database once they have signed an agreement whereby they abide to protecting the anonymity and privacy of the participants under all conditions.

2) Privacy related perceptions among participants: Our participants are all aware of the fact that their data is collected during the study. Such high awareness level, combined with real world exposure to mobile based data collection implies that privacy related aspects can be assessed in an ecologically valid way in the course of the campaign. Thus instead of describing context aware services to naive individuals participating in focus groups, we now have a pool of individuals who are subjects of people centric sensing events for a prolonged time period. Viewed in this way, the campaign months provide an excellent time window for collecting qualitative data concerning privacy related desires and perceptions.

The intention is to engage campaign participants in a range of activities around the topic of privacy, including e.g. individual interviews, focus groups, and participatory design sessions. The goal of the user centric research agenda is to provide the campaign members with the means of ideating privacy friendly context aware services. The qualitative techniques are also expected to yield design guidelines and recommendations concerning privacy in the context of people centric sensing.

III. Early figures about the dataset

The above section describes the methodological aspects related to the Nokia data collection campaign. This section shows the potential of the dataset for driving analyses of various types.

At the time of writing the paper, the campaign population had reached total size of 168 (65 percent males; 35 percent females). The largest age groups present in the campaign were 22-27 and 28-33, with 37 and 30 percent of the sample falling to these age ranges, respectively. The questionnaire data revealed the following occupational range: 63 percent employed, 8 percent not presently employed, 26 percent students, and 3 percent listed their status as 'other'. The sample was relatively advanced technologically. All of the participants had prior experience in using a mobile phone at the time of joining the campaign. 97 percent of the sample categorised themselves as active Internet users.

The total amount of data currently available covers about 715 months of participants' life. If most of the campaign members manage to stay in the campaign for the target of one year, the figure should be close to 2000 months at the end of the campaign. Table III gives an overview of the current amount of each type of data (these numbers can be safely extrapolated for an estimate of the final values). The columns of the table show the absolute number of records of each type and the total recorded time (when applicable)[1].

Data type	Quantity	Duration
GPS points	4,527,539	
Calls	132,109	3,907h
SMS	88,225	
Pictures taken	28,054	
Videos shot	2,163	
Bluetooth scans	15,362,182	
Unique Bluetooth devices	238,221	
WLAN scans	12,568,788	
Unique access points	265,372	
Unique cell towers	46,082	
Accelerometer samples	547,492	1521h
Audio samples	218,021	1817h
Application events	3,569,860	
Unique calendar entries	6,171	
Unique phone book entries	34,053	

TABLE III
Figures about the collected dataset (at the time of writing)

A. Sample composition from a customer centric view

Nokia's customer segmentation model includes 13 segments. The segments are placed along two defining dimensions: low versus high involvement and rational versus aspirational mindset. The involvement axis refers to frequency of replacing one's mobile phone, amount of money invested on the product as well as disposition to use services and functions of the device. The rational vs aspirational axis refers to what consumers expect from the mobile technology. A rational consumer wants to perform variety of tasks with the same mobile phone. He is happy with the mobile phone as long as it performs like he wants it to. An aspirational consumer, on the other hand, would be more likely to use his mobile phone as a fashion accessory. The phone should not look outdated and it is important for an aspirational consumer to keep up to date with the latest fashion trends.

A seven item scale determining the consumer segment of the participant was incorporated in the questionnaire administered during the enrollment session. Figure 3 shows the deviation of the campaign population from the global average, relative to the quadrants in the Nokia segmentation model.

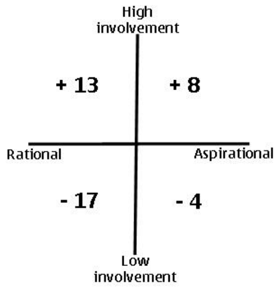

Fig. 3. Deviation (in percentage units) of campaign population from Nokia's global consumer segments

Interestingly, the high involvement quadrants are overrepresented in the present sample pointing to a possible selection bias in the sampling. Since the participants are given high end mobile phones for the duration of the campaign, individuals categorized on the higher side of the involvement axis were perhaps more likely to gravitate toward the study than individuals with a low involvement level.

We were aware of the possible bias in the sampling method but did not engage in countermeasures because of placing high priority on the ability to involve real life social networks in the study. This in, turn, was only possible through deploying an opportunistic sampling method so as to enable clusters of friends, colleagues and family members to participate.

B. Social interaction

As described in Section II-B, the sampling method has been designed so that the social graph between participants is connected. Since we know the phone numbers of the participants, it is easy to track the communications between them. The current data contains a total of 14,042 calls (of a total duration of 288 hours) and 6475 messages between participants. This represents respectively 10.6 and 7.5 percent of the total number of phone calls and SMS's, respectively.

[1] The current version of the client software records only the starting time of calendar entries, but not their end time, which prevents us from estimating the total amount of time covered by them.

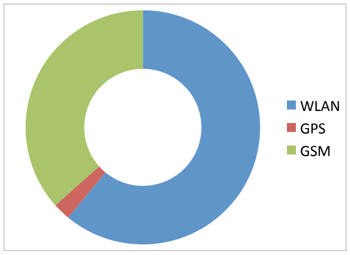

Fig. 4. Relative availability of positioning technologies

Besides calls and messages, Bluetooth is useful in capturing social interaction: Since we forced the phones' Bluetooth visibility on, we can also track when two participants are in range of each other, which means in most cases that they are having some social interaction. The current dataset allows to calculate that an average participant spends 18.6 percent of his time in the range of another participant. Note that WLAN scanning might allow for more accurate estimation of the proximity between participants. Such an approach is currently under development.

C. Positioning

GPS is naturally the most accurate positioning method. As explained in Section II-A, the GPS receiver could not be activated permanently due to its high power consumption. However, the state machine implemented in the data collection software allows to still capture most of the outdoor trajectories of the participants. The fraction of time where participants are collecting detailed GPS data is 2.4 percent.

Besides GPS, less accurate localization techniques can be used, such as WLAN triangulation. By combining GPS data and WLAN scans, we have been able to estimate the location of 57 percent of the access points. Participants are covered by such access points 61 percent of the time, which allows a rough localization. The accuracy of this method depends of course heavily on the density and placement of the access points.

For the remaining fraction of the time, users can simply be positioned using the cell ID information. Note that a shortcoming of the current software is that it reports the ID and received power of only one cell tower at a time, which precludes GSM triangulation. Figure 4 illustrates the time breakdown in terms of positioning methods.

IV. Conclusions

This paper describes a people centric sensing campaign that is currently operating in Lausanne, Switzerland, and run by a corporate research laboratory from the mobile industry. By using a viral recruitment method for participants, the data collected generates insights not only at the personal level but also at the level of social interaction. Early results indicate that the extent to which the campaign has managed to capture social interaction in real life constellations is high. Of the total amount of phone calls and SMSs sent by the campaign participants, 10.6 and 7.5 percent, respectively, have been sent within this population. The Lausanne data collection campaign differs from previously conducted experiments in that the diversity of the campaign population is relatively high: two thirds are employed, while only a quarter of the population are students. The range of sensed aspects is also larger than in previous campaigns: the usage of phones with GPS and WLAN capability enables accurate outdoor and indoor positioning, while the collection of audio and acceleration samples gives precious context information. Finally, the overall size of the data set is unprecedented: the total recorded time will be close to 2000 participant months and the size of the sample is almost 170.

The insights produced by the data remain to be demonstrated by future papers. We anticipate three levels of findings to emerge: theoretical, consumer centric and user centric.

Acknowledgments

The design and setup of the campaign was made when the first author was affiliated with NRC Lausanne. The authors would like to thank the NRC Lausanne team for their outstanding contributions throughout the data collection campaign, and the EPFL Transport and Mobility Laboratory (http://transp-or.epfl.ch/) for their contribution to the data collection client software.

References

[1] E. Miluzzo, N. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. Eisenman, X. Zheng, and A. Campbell, "Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*. ACM, 2008, pp. 337–350.

[2] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabsi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational Social Science," *Science*, vol. 323, pp. 721–723, 2009.

[3] M. Raento, A. Oulasvirta, and N. Eagle, "Smartphones: an emerging tool for social scientists," *Sociological Methods & Research*, vol. 37, no. 3, p. 426, 2009.

[4] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[5] ——, "Eigenbehaviors: Identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1057–1066, 2009.

[6] K. Farrahi and D. Gatica-Perez, "What did you do today?: discovering daily routines from large-scale mobile data," in *Proceeding of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 849–852.

[7] S. Phithakkitnukoon, H. Husna, and R. Dantu, "Behavioral Entropy of a Cellular Phone User," *Social Computing, Behavioral Modeling, and Prediction*, pp. 160–167, 2008.

[8] A. Campbell, S. Eisenman, N. Lane, E. Miluzzo, and R. Peterson, "People-centric urban sensing," in *Proceedings of the 2nd annual international workshop on Wireless internet*. ACM, 2006, p. 18.

[9] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, p. 15274, 2009.

[10] See where your friends are right now. Officia company web site. Google. [Online]. Available: http://www.google.com/latitude

[11] Share your location with maps. Official company web site. Nokia. [Online]. Available: http://blog.ovi.com/2010/01/22/share-your-location-with-maps/

[12] Citysense. Official company web site. Sensenetworks. [Online]. Available: http://www.sensenetworks.com/citysense.php

[13] D. Norman, *The design of everyday things*. Basic Books New York, 2002.