Appeared in:
Intl. Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, 2012

How Many Makes a Crowd? On the Evolution of Learning as a Factor of Community Coverage

Yaniv Altshuler[1], Michael Fire[2], Nadav Aharony[1], Yuval Elovici[2]
and Alex ("Sandy") Pentland[1]

[1]MIT Media Lab, Cambridge
{yanival, nadav, sandy}@media.mit.edu

[2]Deutsche Telekom Laboratories at Ben-Gurion University of the Negev, Israel
{mickyfi, elovici}@bgu.ac.il

Abstract. As truly ubiquitous wearable computers, mobile phones are quickly becoming the primary source for social, behavioral and environmental sensing and data collection. Today's smartphones are equipped with increasingly more sensors and accessible data types that enable the collection of literally dozens of signals related to the phone, its user, and its environment. A great deal of research effort in academia and industry is put into mining this raw data for higher level sense-making, such as understanding user context, inferring social networks, learning individual features, and so on. In many cases, this analysis work is the result of exploratory forays and trial-and-error. In this work we investigate the properties of learning and inferences of real world data collected via mobile phones for different sizes of analyzed networks. In particular, we examine how the ability to predict individual features and social links is incrementally enhanced with the accumulation of additional data. To accomplish this, we use the *Friends and Family* dataset, which contains rich data signals gathered from the smartphones of 130 adult members of a young-family residential community over the course of a year and consequently has become one of the most comprehensive mobile phone datasets gathered in academia to date. Our results show that features such as ethnicity, age and marital status can be detected by analyzing social and behavioral signals. We then investigate how the prediction accuracy is increased when the users sample set grows. Finally, we propose a method for advanced prediction of the maximal learning accuracy possible for the learning task at hand, based on an initial set of measurements. These predictions have practical implications, such as influencing the design of mobile data collection campaigns or evaluating analysis strategies.

Keywords. Sampling size, Social network, Mobile sensing, Inferring attributes

1 Introduction

Mobile phones are increasingly used as social and behavioral data collection instruments. Their increased pervasiveness makes them an ideal wearable sensing platform for location, proximity, communications and context. Eagle and Pentland[1] coined the term "*Reality Mining*" to describe the collection of sensor data pertaining to human social behavior. As the field of computational social science matures, a need for more structured tools and methodologies is further

required. In particular, we are interested in tools that would assist the researcher or practitioner in designing data collection campaigns, understanding the potential of collected datasets and estimating the accuracy limits of current analysis strategy versus alternative ones.

Conducting mobile-phone based field studies is challenging and costly. Some of these costs might include subject compensation, technical system development and maintenance, ongoing support for subject's phone hardware and software and mobile phone plans. There is also the added resource and manpower cost related to constant communication with subject populations, recruitment and minimizing attrition of subjects as well as researchers. On the other hand, such studies give us an unprecedented window into the lives of individuals and entire communities. As demonstrated in [2], conducting user-centric data collection provides a wide range of signals on participants and allows us to directly gather ground-truth and ancillary information such as demographic information, physiological and psychological information, self-perceptions and attitudes and a variety of additional data types gathered via surveys, interviews, and interactions with the subject population. These types of information are usually not available for datasets "donated" to researchers by mobile service providers and other commercial entities.

Mobile-phone studies span along a wide range of studies. Some studies are all encompassing and broadly defined living laboratory or "social observatory" types of studies. These include the Reality Mining study [7], the Social Evolution study [10], and the Friends and Family Study [2]. There are also more targeted studies, focusing on specific research questions or data types, like physical activity levels or environmental impacts.

In these and similar initiatives, there is a continuing and conflictive tension. On one hand, there is an idealistic desire to make the best use of the invested time and resources; collection of as an encompassing a dataset as possible which includes a maximal amount of subject from the target populations. Real-world data is often noisy and challenging to work with. When attempting to establish inferences or generalizations, we wish to increase our confidence in the data by assuring that the dataset is as complete as possible. Indeed, there exist practical considerations of the time and cost of additional resources as well as the fact that due to many reasons, it is never really possible to recruit all members of a target community to any study. However, we ask the question, are all target members really necessary? And how many, exactly, is "enough"?

In this work, we aim to gain a better understanding of the evolution of the process of learning personal features and behavioral properties based on mobile-phone sensed data as we increase the size of the sample group. We investigate issues related to the "coverage" of a given community, i.e., the number of subjects we have access to with respect to the actual size of the measured sub-network. For this analysis, we are less concerned about the specific learned models and their generalizability, and more about using them to study and benchmark the evolution of learning accuracy. Understanding this process is of significant importance to researchers in a variety of fields as it would provide an approximation for the needed level of coverage (sample size vs. community or network size) in order to "learn" specified features for some given accuracy. Alternatively, it could give the investigators an idea of the expected level of accuracy that can be obtained for a given socially-relevant data collection initiative.

To carry out our research we use the *Friends and Family* dataset which contains rich data signals gathered from the smartphones of 140 adult members of a young-family residential community and collected over the course of a year [2], as well as self-reported personal and social-tie information. We build voting classifiers for predicting personal properties such as nationality and religion, based on the participants' SMS messages graph's topology. We demonstrate the characteristics of incremental learning of multiple social and individual properties from raw sensing data collected from mobile phones while the information is accumulated from a multitude of individuals. We observe similar behavior among different learning processes as sample size increases. We also observe a limit, or "saturation", where additional community coverage does not increase the accuracy of prediction, or in some cases, increases it only marginally.

Furthermore, we propose a method for advanced prediction of the maximal learning accuracy possible for the learning task at hand using just the first few measurements. This information can be useful in many ways, including:

- Informing real-time resource allocation for data collection for an ongoing study.
- Estimating the monitoring time needed for desired accuracy levels of a given method.
- Early evaluation of modeling and learning strategies.

The paper is organized as follows: Related work is presented in Section 2. Our analysis is described in Section 3 and concluding remarks in Section 4. Additional technical and methodological details can be found in the Appendix.

2 SCIENTIFC BACKGROUND

In recent years the social sciences have been undergoing a digital revolution heralded by the emerging field of ‘‘Computational Social Science’’. Lazer, Pentland, et al. [3], describe the potential of computational social science to increase our knowledge of individuals, groups, and societies, with an unprecedented breadth, depth, and scale. Computational social science combines the leading techniques from network science[4-6] with new machine learning and pattern recognition tools aimed for the understanding of people's behavior and social interactions [7].

2.1 Mobile Phones As Social Sensors

The pervasiveness of mobile phones the world over has made them a premier data collection tool of choice and they are increasingly used as social and behavioral sensors of location, proximity, communications and context. Eagle and Pentland[1] coined the term "Reality Mining" to describe the collection of sensor data pertaining to human social behavior. They show that by using call records, cellular-tower IDs, and Bluetooth proximity logs collected via mobile phones at the individual level, the subjects' regular patterns in daily activity can be accurately detected[1, 7]. Furthermore, mobile phone records from telecommunications companies have proven to be quite valuable for uncovering human level insights. For example, Gonzales et al. show that cell-tower location information can be used to characterize human mobility and that humans follow simple reproducible mobility patterns[8]. This approach has already expanded beyond academia, as companies like Sense Networks [9] are putting such tools to use in the commercial world to understand customer churn, enhance targeted advertisements, and offer improved personalization and other services.

2.2 Individual Based Data Collection

Data gathered through service providers include information on very large numbers of subjects. However, this information is constrained to a specific domain (email messages, financial transactions, etc.) and there is very little, if any, contextual information on the subjects themselves. The alternative data gathering approach at the individual level allows collection of many more dimensions related to the end user which are many times not available at the operator level.

Madan et al.[10] expand Eagle and Pentland's work[1] and show that mobile social sensing can be used for measuring and predicting the health status of individuals based on mobility and communication patterns, as well as the spread of political opinion within the community[11]. Other examples of using mobile phones for individual-based social sensing are Montoliu et al.[12], Lu et al.[13], and projects coming from CENS center, e.g. Campaignr by Joki et al.[14], as well as additional works described in [15]. Finally, the Friends and Family study, which our paper uses as its data source, is probably the richest mobile phone data collection initiative to

date, in relation to the number of signals collected, study duration and the number of subjects. In addition to mobile phones, there have been other types of wearable sensor-based social data collection initiatives. A notable example is the *Sociometric Badge*, which captures human activity and socialization patterns via a wearable device and is mostly used for data collection in organizational settings [16]. Our results are applicable to these types of studies as well.

2.3 Learning and Prediction of Social and Individual Information.

A great deal of studies involving predicting individual traits and social ties were conducted in the recent years in the general context of social networking. Example include works by Liben-Nowell and Kleinberg [17], Mislove[18] and Rokach et. al. [19], combining machine learning algorithms with social network data in order to build classifiers. In computational learning theory, "Probably Approximately Correct" (PAC) learning tries to solve similar problems of finding efficient classifier functions that depends on the train set sample size [31].

3 ANALYSIS AND RESULTS

The goal of this work is to study and analyze the evolution of the learning process of personal features and behavioral properties with a constant increase in sampling group size. Our analysis focuses less on specific learned models and their generalizability, and is more concerned by their use as benchmark for the learning process as data accumulates. Understanding this process is of significant importance to researchers in a variety of fields as it provides an approximation for the amount of data-samples needed in order to "learn" these features for some given accuracy, or alternatively, revealing the level of accuracy that can be obtained for a given community.

We constructed the SMS messages social network which was created after 65 weeks of the study, using more than 97,000 SMS messages sent between March 1st 2010 and May 31th, 2011. An edge represents users that had at least two common friends in the network (see Figure 4). We then attempted to predict each participants' personal information by using their friends personal information. We divided the social network graph to communities using the *Louvain* algorithm for community detection [23]. We look at the community which the participant belongs to and predict that attribute *a* would be equal to the attribute *a* majority in the community. We then used the aforementioned classifiers on 100 random training groups, each containing a single user (and test groups that contained the rest of the community), using the average performance. We then took 100 random groups of 2 users each, and so on, until testing 100 groups of 90% of the users. Figures 1 and 2 present the results of 5 of the classifiers we have run.

We used the *Gompertz function* in order to model the evolution of this process : $f = ae^{be^{ct}}$

This well known model is flexible enough to fit various social learning mechanisms while providing the following important features:

(a) Sigmoidal advancement; a monotonous increase in accuracy with increase in group size.

(b) The rate at which information is produced is smallest at the start and end of the process.

(c) Asymmetry of the asymptotes, as for any value of *t*, the amount of information gathered in the first *t* time steps is greater than the amount gathered at the last *t* time steps.

The *Gompertz function* is frequently used for modeling a great variety of processes due to its flexibly, such as mobile phone uptake [25], population in a confined space [26] and growth of tumors [27]. In addition, the applicability of the *Gompertz function* for modeling progress of behavior patterns prediction of mobile and social users was demonstrated in [28].

The regression yielded surprisingly accurate results:

(a) **Ethnicity** : Residual standard error: 0.02026, Achieved convergence tolerance: 3.167e-06
(b) **Religion** : Residual standard error: 0.02115, Achieved convergence tolerance: 8.747e-06
(c) **Children** : Residual standard error: 0.01901, Achieved convergence tolerance: 2.464e-06
(d) **Origin** : Residual standard error: 0.02515, Achieved convergence tolerance: 8.127e-06
(e) **Age** : Residual standard error: 0.02056, Achieved convergence tolerance: 6.189e-06

4 CONCLUSION

In this paper we have studied the effects of the amount of sensors information on the ability to predict personal features of community's members. We have shown that the dynamics of this process can be modeled using the *Gompertz* function and hence can be further extrapolated in order to predict bounds on the overall ability to learn specific features. 'This information can be used to inform and design the data collection for an ongoing or future data collection experiment or initiative. This can be done by extrapolating the *Gompertz* function with the regression

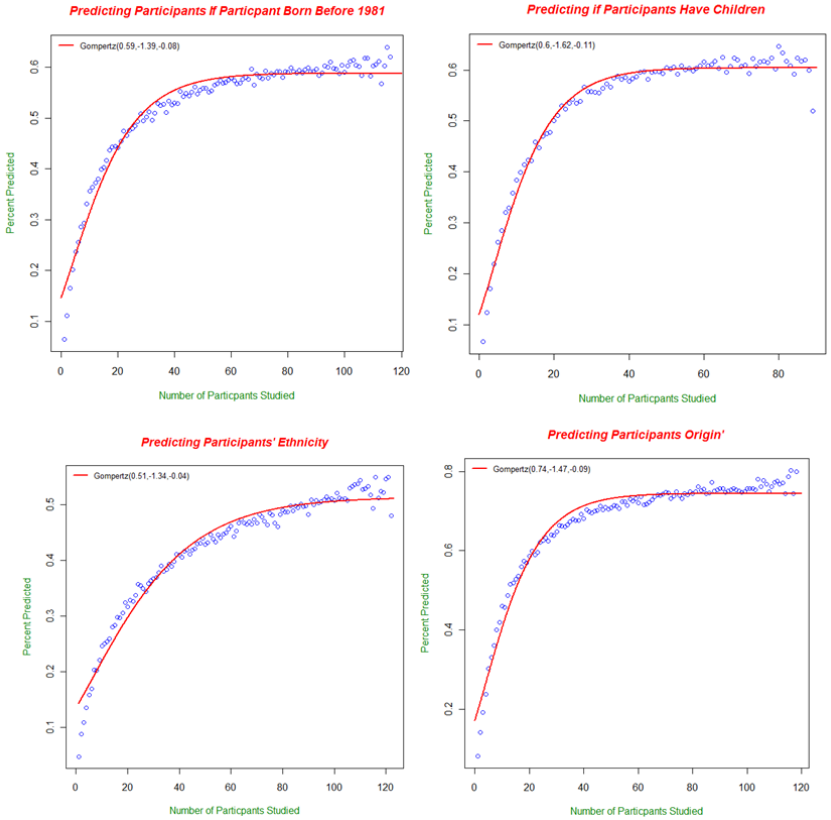

Figure 1. Learning process for predicting participants' origin, ethnicity, age, and parenthood, with data regarding groups of growing sizes. For each group's size, 100 random groups were generated and their performance values averaged. Red lines represent the *Gompertz* regression.

values found in several initial small sampling sets. Based on this extrapolation, an approximation for the maximal amount of information (or accuracy) that can be achieved with large sample sets, as well as the accuracy that a given group size would result in, can be produced.

In addition, correlations between the evolutions of the different learning processes, as depicted in Figure 3, may imply an underlying correlation between the raw data itself, and can hence be used as further validation for correlated features and observations, such as the suggestion that people are more likely to marry within their own ethnic group, as observed in [29,30].

In conclusion, it is also interesting to compare how does the dynamics of the learning process as a function of the community's size correlates with the dynamics of the learning process over time. We intend to conduct this study on the Friends and Family dataset in the coming months.

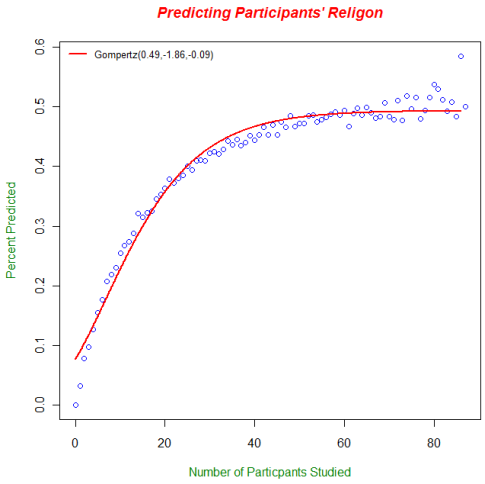

Figure 2. Learning process for predicting participants' religion with data regarding groups of growing sizes. For each group's size, 100 random groups were generated and their performance values averaged. The red line represents the *Gompertz* regression.

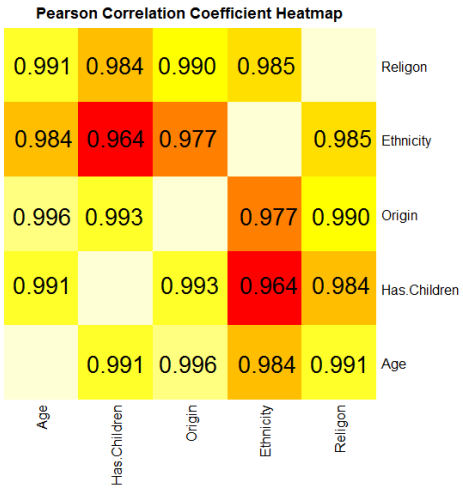

Figure 3. The correlation matrix between the prediction vectors of the 5 classifiers.

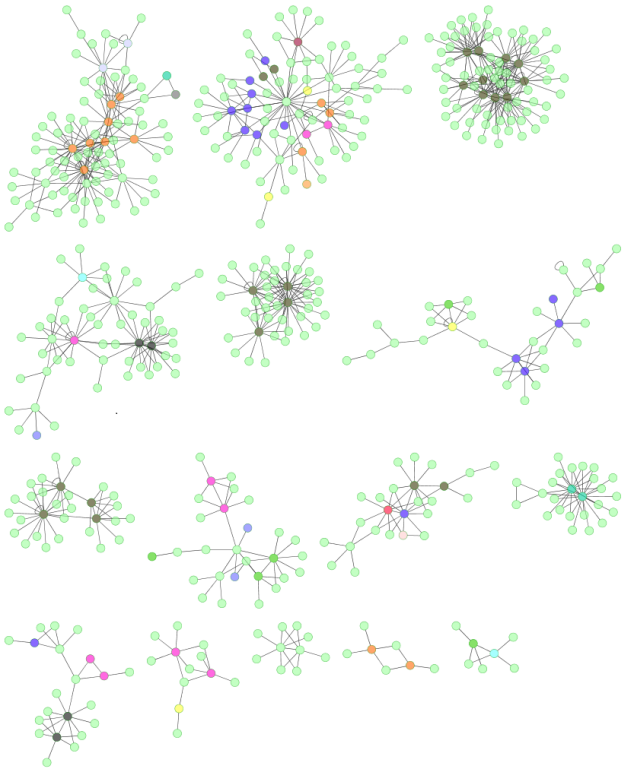

Figure 4. SMS Social Network Graph created over 65 weeks (each unknown node connected to at least two known nodes). Different colors represent different religions.

5 REFERENCES

1. Eagle, N. and A. Pentland, *Reality Mining: Sensing Complex Social Systems.* Personal and Ubiquitous Computing, 2006. **10**: p. 255--268.
2. Aharony, N., et al., *Social fMRI: Investigating and shaping social mechanisms in the real world.* in Pervasive and Mobile Computing, 2011.
3. Lazer, D., et al., *Life in the network: the coming age of computational social science.* Science, New York, NY, 2009. **323**: p. 721.
4. Barabasi and, A.-L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999.
5. Newman, M.E.J., *The structure and function of complex networks.*
6. Watts, D.J. and S.H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature, 1998.
7. Eagle, N., A. Pentland, and D. Lazer, *From the Cover: Inferring friendship network structure by using mobile phone data.* Proceedings of The National Academy of Sciences, 2009. **106**(36): p. 15274-15278.

8. Gonzalez, M.C., A. Hidalgo, and A.-L. Barabasi, *Understanding individual human mobility patterns.* Nature, 2008.
9. Networks., S.; Available from: http://www.sensenetworks.com/.
10. Madan, A., et al., *Social sensing for epidemiological behavior change*, in *Ubiquitous Computing/Handheld and Ubiquitous Computing*. 2010. p. 291-300.
11. Madan, A., K. Farrahi, and D. Gatica-Perez, *Pervasive Sensing to Model Political Opinions in Face-to-Face Networks*. 2011.
12. Montoliu, R. and D. Gatica-Perez, *Discovering human places of interest from multimodal mobile phone data*. 2010. 1-10.
13. Lu, H., et al., *The Jigsaw continuous sensing engine for mobile phone applications*, in *Conference On Embedded Networked Sensor Systems*. 2010. p. 71-84.
14. Joki, A., J.A. Burke, and D. Estrin, *Campaignr: A Framework for Participatory Data Collection on Mobile Phones.* 2007.
15. Abdelzaher, T.F., et al., *Mobiscopes for Human Spaces.* IEEE Pervasive Computing, 2007. **6**(2): p. 20-29.
16. Olguín, D.O., et al., *Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior.* "IEEE Transactions on Systems, Man, and Cybernetics", 2009. **39**(1): p. 43-55.
17. Liben-Nowell, D. and J. Kleinberg, *The link-prediction problem for social networks.* Journal of the American Society for Information Science and Technology, 2007 **58**(7): p. 1019-1031.
18. Mislove, A., et al., *You are who you know: inferring user profiles in online social networks*, in *Web Search and Data Mining*. 2010. p. 251-260.
19. Rokach, L., et al., *Who is going to win the next Association for the Advancement of Artificial Intelligence Fellowship Award? Evaluating researchers by mining bibliographic data.* Journal of the American Society for Information Science and Technology, 2011.
20. Funf. *Funf Project*. Available from: http://funf.media.mit.edu.
21. Hagberg, A.A., D.A. Schult, and P.J. Swart, *Exploring Network Structure, Dynamics, and Function using NetworkX*. 2008.
22. Hall, M., et al., *The WEKA data mining software: an update.* Sigkdd Explorations, 2009. **11**(1): p. 10-18.
23. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of Statistical Mechanics: Theory and Experiment, 2008. **10**.
24. Xie, J. and B.K. Szymanski, *Community Detection Using A Neighborhood Strength Driven Label Propagation Algorithm.* Computing Research Repository, 2011.
25. Rouvinen, P., *Diffusion of digital mobile telephony: Are developing countries different?* Telecommunications Policy, 2006. **30**(1): p. 46-63.
26. Erickson, G.M., *Tyrannosaur Life Tables: An Example of Nonavian Dinosaur Population Biology.* Science, 2006. **313**(5784): p. 213-217.
27. Donofrio, A., *A general framework for modeling tumor-immune system competition and immunotherapy: Mathematical analysis and biomedical inferences.* Physica D-nonlinear Phenomena, 2005. **208**(3-4): p. 220-235.
28. Pan, W., N. Aharony, and A. Pentland, *Composite Social Network for Predicting Mobile Apps Installation*, in *Intelligence, AAAI-11* 2011: San Francisco, CA, 2011.
29. Kalmijn, M., *Intermarriage and Homogamy: Causes, Patterns, Trends.* Annual Review of Sociology, 1998. **24**(1): p. 395-421.
30. McPherson, M., L. Smith-Lovin, and J.M. Cook, *Birds of a Feather: Homophily in Social Networks.* Annual Review of Sociology, 2001. **27**(1): p. 415-444.
31. Haussler D., Part 1: Overview of the Probably Approximately Correct (PAC) learning framework. 1995.

APPENDIX

A. METHODOLOGY

A.1. Mobile Data Collection System

Aharony et al.[2] developed a social and behavioral sensing platform that runs on Android operating-system based mobile phones which can continuously record a broad range of data signals. Each type of signal collected by the system is encapsulated as a conceptual ‘‘probe’’ object. The ‘‘probes’’ terminology is used rather than ‘‘sensors’’, as probes include traditional sensors such as GPS or accelerometer, as well as other types of information not traditionally considered as sensor data, like file system scans or logging user behavior inside applications. Additional signals include information such as cell tower ID, wireless LAN IDs, proximity to nearby phones and other Bluetooth devices, call and SMS logs, statistics on installed phone applications, running applications, media files, general phone usage and other accessible information.

The dataset described in the next section was collected using this system with a configuration that included over 25 different types of data signals. The deployment also included an on-phone survey component and integrated applications such as an alarm clock app. Figure 5 illustrates the deployed system configuration, enabling automated data upload, as well as remote configuration settings and remote updating of the system itself. Figure 6 gives an overview of the back-end side of the system. The software system, named “*Funf*”, has been released as an open source and available at [20].

The “Friends and Family” living laboratory study was conducted over a period of 15 months between March 2010 and June 2011 with a subject pool of 140 individuals. It is the first study conducted under the Social fMRI methodology which uses mobile phones together with a data-rich collection approach to create a “virtual imaging chamber” around a community *in-situ*[2]. To the best of our knowledge, it is the most comprehensive mobile phone experiment performed in academia to date.

A.2. Community Overview.

The research goals of the Friends and Family study touch on many aspects of life, from better understanding of social dynamics, to health, to purchasing behavior and to community organization. It was conducted with members of young-family residential living community adjacent to MIT. All members of the community were couples and at least one of the members affiliated with the university. The community was composed of over 400 residents, approximately half of which had children. In March 2010, the first pilot phase of the study was launched with 55 participants, and in September 2010, the second phase of this study was launched with 85 additional participants. The participants were selected randomly and in a way that would achieve a representative sample of the community and sub-communities.

A.3. Privacy Considerations.

The study was approved by the Institutional Review Board (IRB) and conducted under strict protocol guidelines. One of the key concerns in the design of the study was the protection of participant privacy and sensitive information. For example, data is linked to coded identifiers

for participants and not their real world personal identifiers. All human-readable text, like phone numbers and text messages, are captured as hashed identifiers and never saved in clear text. Collected data is physically secured and de-identified before being used for aggregate analysis.

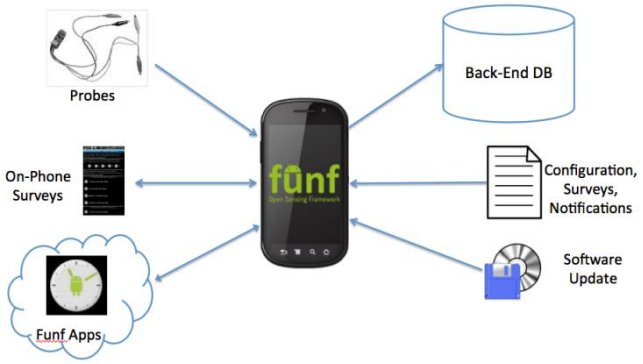

Figure 5. Friends and Family Phone System Overview

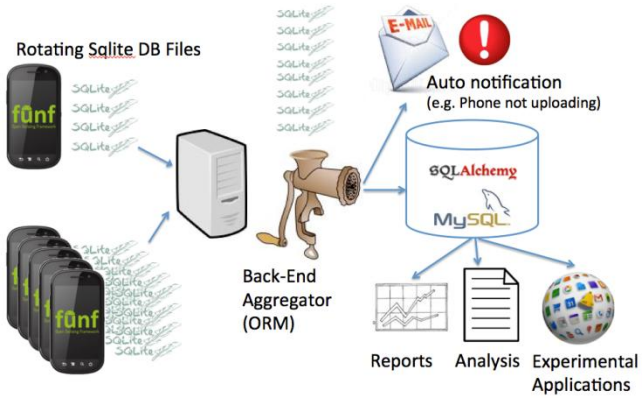

Figure 6. Back-End Data Aggregation Overview

A.4. Friends and Family Dataset.

To the best of our knowledge, the dataset generated from the study is probably the largest and richest dataset ever collected on a residential community to date. The accumulated size of the database files uploaded from the study devices adds up to over 60 Gigabytes. The data is composed of over 30 million individual scan events (for all signals combined), where some events capture multiple data signals. As an example, the dataset includes:

- 20 million WiFi scans which, in turn, accumulated 243 million total scanned device records.
- 5 million Bluetooth proximity scans which, in turn, accumulated 16 million total scanned device records.
- 200,000 phone calls.
- 100,000 text messages (SMS).

In the current analysis presented in this paper, we give special focus to the data that was collected in November 2010 and April 2011 after which the mobile platform was improved, new features, such as different call types where added, and several hardware problems where fixed. These two months had no major holiday breaks in the academic schedule of the university and the bulk of participants were physically on campus.

In addition to the phone-based data, the study also collected personal information on each of the participants. The dataset includes information on age, gender, religion, origin, current and previous income status, ethnicity, and marriage information, among others.

A.6. Social Network Predictions

Another method for predicting a participant's personal information details is using the participants' different social networks. Using the data collected in the study we can span different types of social networks between the participants according to different interaction modalities. Namely, we can define the following social networks:

- **SMS Social Network**: we can construct the user SMS messages social network (see Figure 4) as weighted graph $G_s =< V_s, E_s >$ according to the SMS messages the participants sent. Each weighted link $e = (u, v, w) \in E_s$ in this social network represents a connection between two different phone numbers $u, v \in V$, while w is the strength of the link defined as the number of SMS message send between the two phone numbers[1].
- **Bluetooth Social Network:** we can construct the social network weighted graph $G_s =< V_B, E_B >$ of face-to-face interaction according to information collected about nearby Bluetooth devices. Each link $(u, v, w) \in E_s$ in this social network represent the fact that the two devices $u, v \in V_B$ encounter each other at least one time, while the w is the strength of the link defined as the number of times the two devices encounter each other.
- **Calls Social Network:** Similar to the SMS social network, we can construct calls social network $G_C =< V_C, E_C >$ according to the participants` phone calls. In this social network each link $(u, v, w) \in E_C$ represent different call that was made between two different phone numbers ,$u, v \in V_C$, while w is the strength of the link defined as the number of calls between *u and v*.

By using the above defined social network, together with different graph theory algorithms, we can predict different types of personal and social information. In order to predict the participants' ethnicity, we used the SMS social network (Figure 4). We used the *Louvain* algorithm for community detection[23], which separates the graph into disjoint groups.

Assuming that at each iteration we have information on the ethnicity of at least some of the nodes, we generate an ethnicity prediction for the members of each detected community based on the ethnicity of the majority of known nodes in that community. This is similar to the ideas of the label propagation approach[24] and in [18].

---

[1] In some cases the number interaction may not be accurate due to the fact we do not have the full connection information on phone number outside the study