Trade-offs in Social and Behavioral Modeling in Mobile Networks

Yaniv Altshuler[1], Michael Fire[2], Nadav Aharony[1], Zeev Volkovich[3], Yuval Elovici[2],
and Alex (Sandy) Pentland[1]

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

[2] Deutsche Telekom Lab, Department of Information Systems Engineering
Ben-Gurion University
{mickyfi,elovici}@bgu.ac.il

[3] Department of Software Engineering
Ort Braude College
vlvolkov@braude.ac.il

Abstract. 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 behavior prediction. In this work we investigate the properties of learning and inferences of real world data collected via mobile phones. In particular, we look at the dynamic learning process over time with various sizes of sampling groups and examine the interplay between these two parameters. We validate our model using extensive simulations carried out using the "*Friends and Family*" dataset which contains rich data signals gathered from the smartphones of 140 adult members of a young-family residential community for over a year and is one of the most comprehensive mobile phone datasets gathered in academia to date.

Keywords: Machine Learning, Social Networks, Mobile Networks

1. Introduction

Mobile phones, and increasingly smartphones, have become an integral part of many people's everyday lives. Users carry their smartphone almost everywhere and use it in order to perform many of their day-to-day communication and activities.
The pervasiveness of mobile phones has made them popular scientific data collection tools, as social and behavioral sensors of location, proximity, communications and context. Eagle and Pentland [1] coined the term ``Reality Mining" to describe collections of sensor data pertaining to human social behavior. While existing work has demonstrated results for modeling and inference of social network structure and personal information out of mobile phone data, most are still mainly proofs of concept in a nascent field. The work of the "data scientist" is still that of an artisan, using personal experience, insight, and sometimes a "gut feeling" in order to extract meaning out of the plethora of data and noise.

As the field of computational social science matures, there is need for more structured methodology 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 vs. alternative ones. Such a methodology would facilitate the process of maturing from a field of "craft" into a field of science.

In this work, we present a first step in this direction. Specifically, we investigate the learning and prediction of social and individual models from raw phone-sensed data. We focus on social ties and individual descriptors that can be tied to social affiliation and affinity. For these prediction tasks, we look at the trade-off between the time period the data is collected in and the number of people that form the sample group. To do this, 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 for over a year[2], in addition to self-reported personal and social-tie information. Preliminary results were described in [37] and [38].

We first build classifiers for predicting personal properties such as nationality or gender. We then proceeded to predict more complicated social links such as the subject's life-partner or "significant other". We demonstrated characteristics of the incremental learning of multiple social and individual properties from raw sensing data collected from mobile phones, as the information is accumulated over time, or alternatively, as the sample size is increased. We study the interplay between the change in time and the growth in sample size and present preliminary results that indicate that such a trade-off exists and that it reflects the network effect of the domain.

2. Related Work

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 sciences to increase our knowledge of individuals and groups with an unprecedented breadth, depth, and scale. Computational social sciences combine the leading techniques from network sciences [4-6] with new machine learning and pattern recognition tools specialized 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 in uncovering human level insights. For example, Gonzales et al. illustrate how cell-tower location information can be used to characterize human mobility, based on the observation that humans follow simple reproducible mobility patterns[8]. This approach has already expanded beyond academia with companies like Sense Networks [9] employing such tools in the commercial world to understand customer churn, enhance targeted advertisements, offer improved personalization and many other services.

2.2. Individual Based Data Collection

On one hand, data gathered through service providers includes information on a very large numbers of subjects, but on the other hand, 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 approach of gathering data 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] follow up on Eagle and Pentland's work [1] and show that mobile social sensing can be used measure and predict the health status of individuals based on mobility and communication patterns. They also investigate the spread of political opinion within a community [11]. Other examples for using mobile phones for individual-based social sensing are those by Montoliu et al. [12], Lu et al. [13], and projects coming out of the CENS center, e.g., Campaignr by Joki et al. [14] as well as additional works as 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 with regards to the number of signals collected, study duration, and the number of subjects. The technical advancements in mobile phone platforms and the availability of mobile software development kits (SDKs) to any developer makes the collection of Reality Mining types of data easier to collect than ever before. Analysis of the security aspects of this trend can be found in [36].

In addition to mobile phones, a notable example for wearable sensor-based social data collection initiative is the *Sociometric Badge* by Olguin et al., capturing human activity and socialization patterns via a wearable sensor badge, that has been used mostly for in organizational settings [16]. The results of our work are applicable to these types of studies as well.

2.3. Learning and Prediction of Social and Individual Information

Many studies which involve predicting individual traits and social ties have been conducted in the recent years within the general context of social networking. Relevant works have been published 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.

3. Data Collection and Analysis Methods

3.1. Friends and Family Dataset

To the best of our knowledge, the dataset generated from this 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 phone 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. Just as 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 analysis presented in this paper we give special focus to the data that was collected in November 2010 and April 2011, after the mobile platform was improved, new features, such as different call types where added, and several hardware problems where fixed. These two months were without a major holiday break 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 contained personal information on each participant, such as age, gender, religion, origin, current and previous income status, ethnicity, marriage information, and more.

3.2. Machine Learning Predictions

In order to evaluate learning over time, which is the main goal of our current work, we needed a set of learning and prediction models to work with. These are mostly illustrative models which enable us to conduct our main analysis. In order to achieve our final goal of predicting participants' personal and social information, we utilized two approaches; a machine learning approach, described in this section and a social network based prediction approach, described in the following section.
The first step in applying the machine learning methodology is to create feature vectors for each participant in the study. Each feature vector contains information on the participant's communication and phone usage patterns as they were collected during the study.

In order to cope with the huge amount of data collected during the study, we developed a code using C# and *Python's NetworkX* library [21]. Our code parsed the collected data and extracted feature vectors for each participant. We extracted 32 different features within a specified time interval. Namely, we collected the following features for each participant:

- **Internet usage features**: we calculated the number of distinct searches performed using the phone's browser and the number distinct bookmarks saved by the user.
- **Calls pattern features**: we computed the total number of calls, the number of unique phone numbers each user was in contact with and the total duration of all calls. We had also calculated the number of incoming/outgoing/missed calls and the total call duration, per call type.
- **SMS messages pattern features**: we computed the total number of SMS messages, the number of unique phone numbers each participant connected with via SMS and the of total incoming/outgoing SMS messages.
- **Phone applications related features:** we counted the number of applications installed and uninstalled on each device. We also computed the total number of currently running applications (originally sampled every 30 seconds).
- **Alarm features:** we counted the number of alarm-clock alarms and the number of "snooze" presses for each participant that used our alarm clock app.
- **Location features:** we calculated the number of different cellular cell tower ids and the number of different wifi network names seen by the smartphone**.** These features act as a rough indication of the number of different locations a participant visited during the time period.

Our next step was to extract all participant features for different time intervals. Using the extracted features, we can build different classifiers that are able to predict the participants` personal information. We used the *WEKA* software [22] in order to test different machine learning algorithms. In our experiments, we evaluated a number of popular learning methods: *WEKA*'s C4.5 decision trees, Naive-Bayes, Rotation-Forest, Random-Forest, and AdaBoostM1. Each classifier was evaluated using the 10-fold cross validation approach and in order to compare results between different classification algorithms, we used each classifier's Area Under Curve or AUC measure (also referred to as ROC Area) and F-measure results. In order to obtain an indication of the usefulness of various features, we analyzed their importance using *WEKA*'s information gain attribute selection algorithm.

Using the machine learning approach we built five different classifiers which predict the following: (1) the gender of the participant, (2) whether the participant is a student or not, (3) whether the participant has children or not, (4) whether the participant is above the age of 30, and (5) whether the participant is a native US citizen or not.

3.3. Social Network Predictions

Another method for predicting a participant's personal information details is by 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 constructed the community's SMS messages social network as a 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 connections between two different phone numbers $\mathrm{u}, \mathrm{v} \in \mathrm{V}$, while w is the strength of the link defined as the number of SMS message sent between the two phone numbers.
- **Bluetooth Social Network:** we constructed a weighted network graph $G_s =< V_B, E_B >$ of face-to-face interactions 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 $u, v \in V_B$ encountered each other at least once, while w is the strength of the link, defined as the number of times the two devices meet.
- **Calls Social Network:** similar to the SMS social network, we can construct a network based on the participant's call graph $G_C =< V_C, E_C >$ according to the participants` phone calls. In this social network, each link $(u, v, w) \in E_C$ represents the fact that at least one call 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 social networks defined above, together with different graph theory algorithms, we can predict different types of personal and social information. In order to predict the participants' significant other we analyzed the Bluetooth social network. We predicted that each participant's significant other is the person that the participant spent the most time with during the measured interval. Namely, let $u \in V_B$ then:

$$\textit{significant-other(u))}= \{v | (u, v, w) \in E_B \textit{ and } \forall (u`, v`, w`) \in E_b \; w > w`\}$$

In order to predict the subjects' ethnicity we used the SMS social network, using the *Louvain* algorithm for community detection [23], which separates the graph into disjoint groups.

For each iteration we assumed that we had information on the ethnicity of at least some of the nodes. We generated an ethnicity prediction for the members of each detected community based on the ethnicity of the majority of known nodes in that community (see more details in [24] or [18]).

3.4. Prediction Accuracy Evolution over Time

The goal of this work is to study and analyze the trade-off between the increased time given for the learning process of personal features and behavioral properties, and an increase in the sample size. For this analysis, we worry less about the specific learned models and their generalizability, and more about using them to study and benchmark the evolution of the learning process as the data accumulates. Understanding this process is of significant importance to researchers in a variety of fields, as it would provide an approximation of the amount of time that is needed in order to "learn" these features for some given accuracy, or alternatively, the level of accuracy that can be obtained for a given duration of time.

To generate the models presented here, we collected the performance results of the system using many combinations of learning time X sample sizes. Each predictor/classifier was executed on data gathered between November 1th and November 30th, 2010. Starting from an input of a single day (November 1st), in each consecutive execution, another day of data was added to the input so that iteration #1 was on data from November 1st, execution #2 had input of data for two days, November 1 and 2 together, and so on until the accumulation of 30 days in which the classifier ran on data from the entire month of November. This process was repeated for varying sizes of the community, starting from 2% of the users to 90%.

4. Results

Following are several examples of the evolution of the learning process of personal and social traits from mobile phone over time (namely, for growing segments of time used for collecting the learned data):

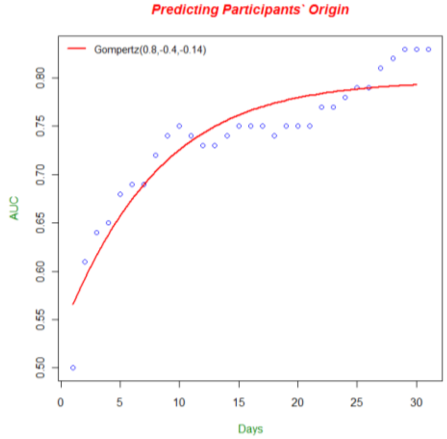

Figure 1. Participants' origin Naïve-Bayes classifiers AUC results.

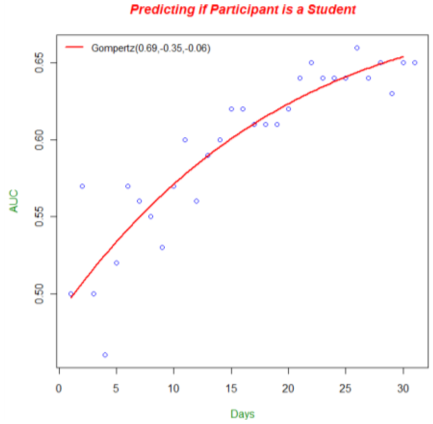

Figure 2. Predicting if the Participant is a student over time: Rotation-Forest Classifier AUC results.

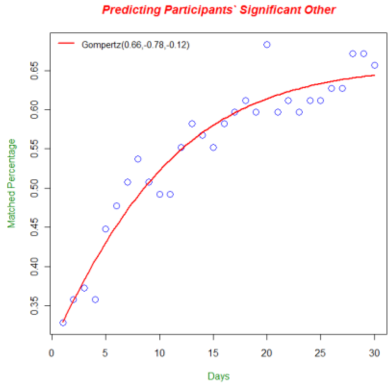

Figure 3. Predicting significant other over time (the node with the maximum strength).

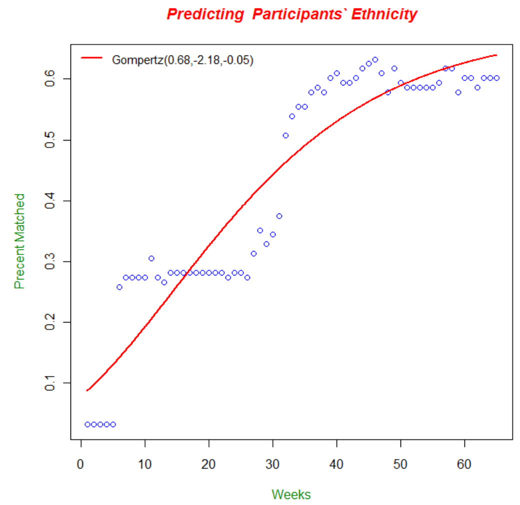

Figure 4. Predicting ethnicity using SMS social network over time (65 weeks) using the Louvain Algorithm.

The following charts illustrate the improvement in the prediction performance, as a function of the sample size (namely, the number of people whose data is used for the learning process) :

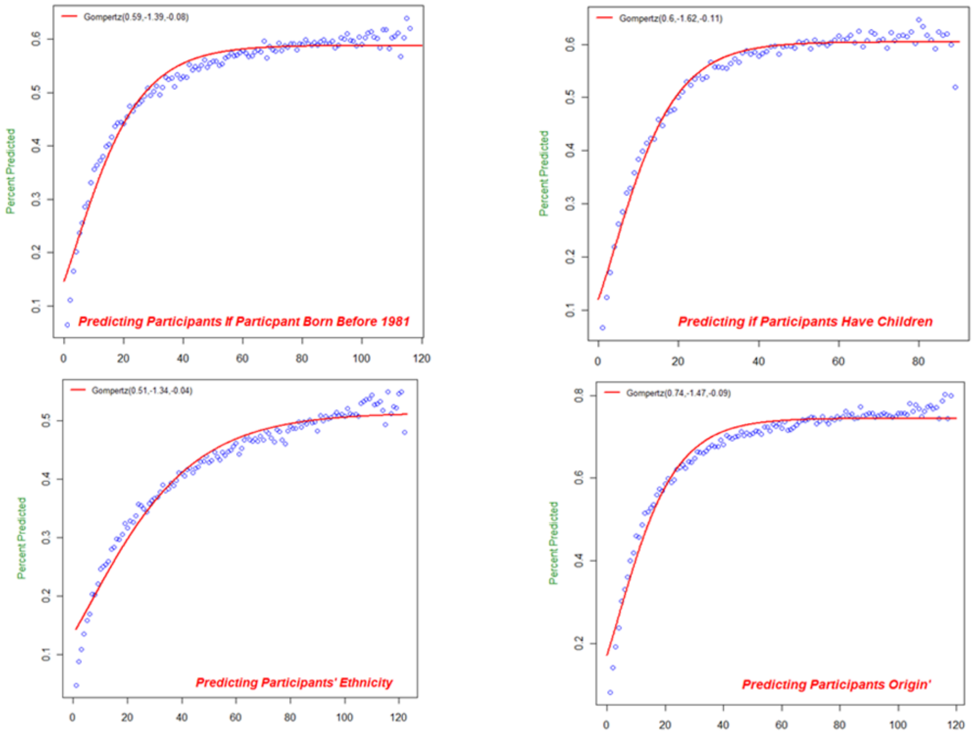

Figure 5. The evolution of the learning process with the growth of the sample size (number of users as the X axes).

After examining the way each of the properties affects the learning process separately, we can now examine the trade-off between an increase in the learning time and an increased sample size. The following charts represent the performance of the learning process (heights) as a function of the number of people of the sampling size (the longer axes) and the time in days. Notice that using the following charts the allocation of data collection and analysis resources can be empirically optimized. For example, given a data collection experiment that has X participants and that is going to last T days, the gradient of the appropriate *f(X,T)* (corresponding to the trait that is to be analyzed) can be numerically calculated, hinting on the best use of any additional resources (i.e. shall we recruit additional participants, or increase the duration of the experiment).

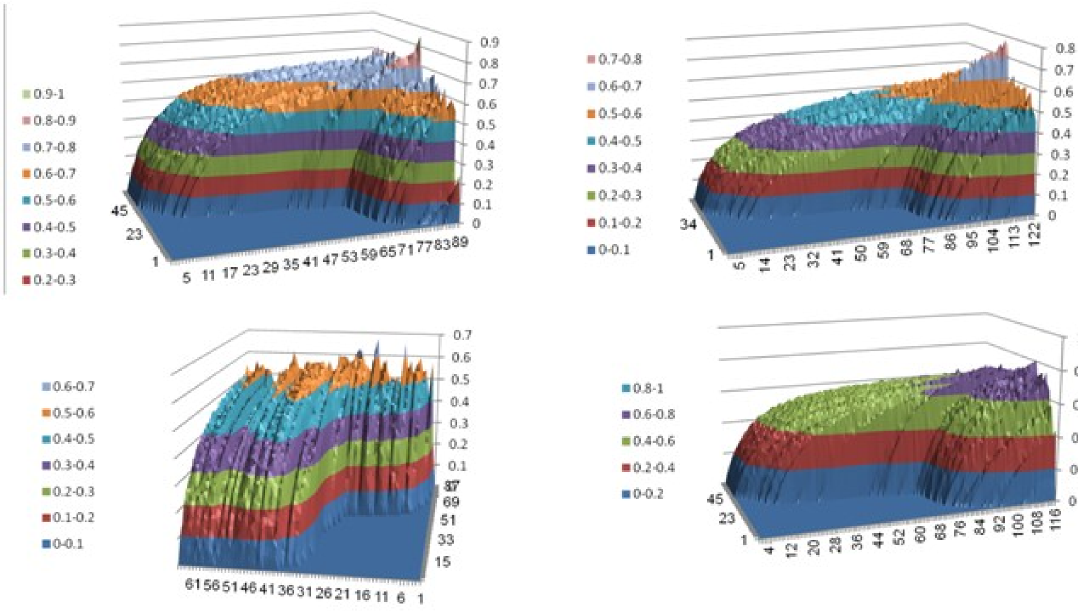

Figure6. An illustration of the trade-off in the learning performance of a mobile social network (height of the function) between the time axis (from 1 to 61 days) and the number of people in a sample group (from 1 to 120). These four charts correspond to the learning of age (top-left), parenting (top-right), ethnicity (bottom-left) and religion (bottom-right).

5. Conclusions

While there will always be the need for the expert and experienced “data artisan”, the exponential increase in accumulated data and the rise of a big-data ecosystem creates an imperative need to design more accurate science and engineering of data collection, processing, and analysis. Our work is a building block towards this larger goal.

In this work we have discussed the trade-off between increased time (and data over time) and increased sample size, regarding the learning of personal features of mobile network users. For this, we have studied a comprehensive dataset of a mobile network containing phone calls, SMSs, and web activity, as well as self-reporting friendship queries.

We demonstrated the characteristics of incremental learning of multiple social and individual properties from raw sensing data collected from mobile phones as the information is accumulated over time. We then did the same for a systemic increase in the sample size used by the learning process. This was done through the use of state of the art techniques for machine learning using all the possible scenarios in terms of learning time and sample size.

We have presented the results of this analysis that hints to an inherent trade-off which is likely to be dominated by the "network-effect" of the domain the data is taken from. In the future stages of this research we intend to develop a mathematical model for the evolution of the learning process (with time and sample size) as well as suggest a mathematical model for the correlation and trade-off between the two.

In addition, it would be interesting to examine the correlation between the efficiency of the learning process and the way we select the identity of the network members whose activities we monitor. Crafting an optimal deployment scheme for monitoring agents throughout the network might provide an additional increase in the convergence rate of the learning mechanism, significantly enhancing our ability to achieve reliable predictions using a given amount of network access. For example, should we aim for symmetric deployment of our monitoring resources? Preliminary results [39] hint though that this might not necessarily be the case.

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. Barabasiand, 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. abs/1105.3.
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. Krishnamurthy, B. and C.E. Wills, On the leakage of personally identifiable information via online social networks. *Computer Communication Review*, 2009. 40(1): p. 7-12.
30. Binde, B.E., R. McRee, and T.J. O`Connor, Assessing Outbound Traffic to Uncover Advanced Persistent Threat. 2011, *Sans Institute*.
31. *Solutionary*, White Paper: The Advanced Persistent Threat (APT). 2011.
32. Brunner, M., et al., Infiltrating Critical Infrastructures with Next-Generation Attacks. 2010, Fraunhofer-Institute for Secure Information Technology SIT Munich.
33. Kalmijn, M., Intermarriage and Homogamy: Causes, Patterns, Trends. *Annual Review of Sociology*, 1998. 24(1): p. 395-421.
34. 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.

35. Dey, A.K., et al., Getting Closer: an Empirical Investigation of the Proximity of Users to Their Smart Phones. *In Proc. of the 13th international conference on Ubiquitous computing* (2011), p. 163-172.
36. Altshuler, Y., Aharony, N., Elovici, Y., Pentland A., and Cebrian, M., Stealing Reality: When Criminals Become Data Scientists (or Vice Versa). *In IEEE Intelligent Systems* (2011), 26(6), p. 22-30.
37. Altshuler, Y, Fire, M., Aharony, N., Elovici, Y., and Pentland A., How Many Makes a Crowd? On the Correlation between Groups' Size and the Accuracy of Modeling*, SBP*, 2012
38. Altshuler, Y, Fire, M., Aharony, N., Elovici, Y., and Pentland A., Incremental Learning with Accuracy Prediction of Social and Individual Properties from Mobile-Phone Data*, Arxiv preprint* arXiv:1111.4645, 2011
39. Altshuler, Y., Wagner, I.A. and Bruckstein, A.M., On Swarm Optimality In Dynamic And Symmetric Environments, *Economics*, 7, 11-18, 2008