Incremental Learning with Accuracy Prediction of Social and Individual Properties from Mobile-Phone Data

Yaniv Altshuler
MIT Media Lab
77 Mass. Ave.
Cambridge, MA 02139
Email: yanival@media.mit.edu

Nadav Aharony
MIT Media Lab
77 Mass. Ave.
Cambridge, MA 02139
Email: nadav@media.mit.edu

Michael Fire
Telekom Innovation Labs
Ben Gurion University
Beer-Sheva 84105, Israel
Email: mickyfi@bgu.ac.il

Yuval Elovici
Telekom Innovation Labs
Ben Gurion University
Beer-Sheva 84105, Israel
Email: elovici@bgu.ac.il

Alex "Sandy" Pentland
MIT Media Lab
77 Mass. Ave.
Cambridge, MA 02139
Email: sandy@media.mit.edu

***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 regarding the phone, its user, and their environment. A great deal of research effort in academia and industry is put into mining this 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. Adding to the challenge, the devices themselves are limited platforms, hence data collection campaign must be carefully designed in order to collect the signals in the appropriate frequency, avoiding the exhausting the the device's limited battery and processing power. Currently however, there is no structured methodology for the design of mobile data collection and analysis initiatives. In this work we investigate the properties of learning and inference of real world data collected via mobile phones over time. In particular, we analyze how the ability to predict individual parameters and social links is incrementally enhanced with the accumulation of additional data. To do so we use the *Friends and Family* dataset, containing rich data signals gathered from the smartphones of 140 adult members of an MIT based young-family residential community for over a year, and is one of the most comprehensive mobile phone datasets gathered in academia to date. We develop several models for predicting social and individual properties from sensed mobile phone data over time, including detection of life-partners, ethnicity, and whether a person is a student or not. Finally, we propose a method for predicting the maximal learning accuracy possible for the learning task at hand, based on an initial set of measurements. This has various practical implications, such as better design of mobile data collection campaigns, or evaluating of planned analysis strategies.**

I. Introduction

Smartphones have become an integral part of many peoples everyday lives. Users carry their phone almost everywhere, using it as their main access point for many of their day-to-day activities. These include connecting with family and friends via voice calls or text messaging, searching for information on the Internet, installing and using different mobile applications for business and leisure, using various location based services such as navigation instructions, or simply using the smartphone as an alarm clock.

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 the collection of sensor data pertaining to human social behavior. While existing works have 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 "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 a more structured methodology, to assist researchers in designing data collection campaigns, predicting the potential of collected data, and estimating the accuracy limits of the analysis. Such a methodology would facilitate the process of maturing from a field of craft into a field of science and engineering.

In this work, we present a first step in this direction, investigating 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. We examine the dynamics of the learning process over time, analyzing how the ability to predict individual parameters and social links is enhanced with the accumulation of additional data.

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], as well as self-reported personal and social-tie information. We first build classifiers for predicting personal properties like nationality or gender. We then proceed to predict more complicated social links such as the subjects life-partner, or "significant other".

We then show that the improvement in social prediction accuracy over time can be modeled using the *Gompertz* function - a known mathematical model that has been used to approximate many processes in a variety of fields, including growth of tumors and adoption of technological services in communities, among others. Using this insight we propose a novel method for a-priori prediction of the maximal learning accuracy possible for the learning task at hand, using just the first few measurements. This method can be used for efficient real-time resource allocation for data collection, ongoing data collection campaign, as well as estimating accuracy limits and time needed for desired accuracy level of a given method.

The paper is organized as follows: We start by presenting related work in Section II. In section III we discuss the methodology of the experiment and our learning techniques. Section IV contains the results, and discussion and concluding remarks are given in Section V.

II. Scientific Background

In recent years the social sciences have been undergoing a digital revolution, heralded by the emerging field of "computational social science", having the potential to increase our knowledge of individuals, groups, and societies, with an unprecedented breadth, depth, and scale [3]. This field combines the leading techniques from network science [4]–[6] with new machine learning and pattern recognition tools specialized for the understanding of people's behavior and social interactions [1], [7].

Using call records, cellular-tower IDs, and Bluetooth proximity logs, collected via mobile phones at the individual level, the subjects' social network can be accurately detected, as well as regular patterns in daily activity [8]. Mobile phone records from telecos have proven to enable uncovering of human level insights: cell-tower location information were used to characterize human mobility [9], socioeconomic status [10], and even health [11]. This approach has expanded beyond academia, as companies are putting such tools to use in the commercial world to understand customer churn, enhance targeted advertisements, and offer improved personalization and other services [12].

Although data owned by service providers contains information on very large numbers of subjects, this information is constrained to specific domains (email messages, financial transactions, etc.), and has very little, if any, contextual information on the subjects themselves. Data collection at the individual level, on the other hand, allows collecting many more dimensions related to the end user, many times not available at the operator level. Madan et al. had shown that mobile social sensing can be used for measuring and predicting the health status of individuals based on mobility and communication patterns [11]. They had also investigated the spread of political opinion within a community [13]. Other examples for using mobile phones for individual-based social sensing are those by Montoliu et al. [14], Lu et al. [15], and projects coming from CENS center [16], and additional works as described in [17].

The technical advancements in mobile phone platforms and the availability of software development kits to any developer is making the collection of *Reality Mining* type of data easier than ever before. 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 by Olguin et al. which captures human activity and socialization patterns and are used mostly for data collection in organizational settings [18]. Additional works that focus on methods for learning and prediction of social and individual properties from mobile phone data can be found in [19]–[21].

III. Materials and Method

A. Methodology

We evaluate our model 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, as well as self-reported personal and social-tie information [2].

Based on data collected from these networks, we have developed *classifiers* capable of accurately predicting personal, behavioral and social attributes of the network's users. We then studied the correlation between the amount of *time* an attacking agent monitors its host victim, and the accuracy of the information it produces, using our classifiers. We show that this process can be modeled using a *Gompertz function*. Based on this insight, we show how the accuracy of inferring personal and social features from mobile data can easily be approximated using an extrapolation of this Gompertz based model.

B. Data Collection Platform

Data was collected using our proprietary *Android* based platform [2]. Monitored signals included traditional sensors such as GPS or accelerometer, file system scans as well as user behavior patterns inside third party applications. Other monitored signals were cell tower ID, wireless LAN IDs; proximity to nearby phones and Bluetooth devices; call and SMS logs; statistics on installed phone applications, running applications, media files, general phone usage; and other accessible information.

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.

To the best of our knowledge, the dataset generated in the study is among the largest and richest ever collected on a residential community to date. The dataset contains 20 million WiFi scans (243 million scanned devices), 5 million

Bluetooth proximity scans (16 million scanned devices), over 200,000 phone calls, 100,000 text messages, and more. The study also collected self-reported personal information on each participant, such as age, gender, religion, origin, current and previous income status, ethnicity, and marital status.

C. Community Overview

The experiment was conducted with members of young-family residential living community adjacent to MIT. All members of the community are couples, and at least one of the members is affiliated with the university. The community is composed of over 400 residents, approximately half of which have 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, in a way that would achieve a representative sample of the community and sub-communities.

D. 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. Coded identifiers for participants were used, and all human-readable text was hashed before streaming. Collected data was physically secured and de-identified before being used for aggregate analysis.

E. Building the Classifiers

We created feature vectors for each participant in the study, containing information on the participant's communication and phone usage patterns. We extracted the following 20 different features for each participant:

- **Internet usage features**: number of searches performed using the phone's browser; number of bookmarks saved.
- **Calls pattern features**: number of incoming / outgoing / missed calls; number of unique phone numbers per call type; total duration of calls.
- **SMS messages pattern features**: number of incoming / outgoing SMS messages; number of unique phone numbers per SMS type.
- **Phone applications related features**: number of applications installed and uninstalled; total number of currently running applications (sampled every 30 seconds).
- **Alarm features**: number of alarm-clock alarms; number of "snooze" presses.
- **Location features**: number of different cell tower IDs; number of different WiFi network names seen by the phone.

We used WEKA [22] for implementing popular machine learning algorithms and selecting the best classifier for each attribute. Specifically, we tested *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, using each classifier's Area Under Curve (AUC) measure and F-measure, and analyzed using WEKA's *Information Gain Attribute Selection Algorithm*.

IV. Prediction Accuracy Evolution over Time

Each classifier was executed on data gathered between Nov. 1^{st} and Nov. 30^{th}, 2010. Starting from an input of a single day, in each execution the consecutive day of data was added to the input, (so that iteration #1 was on data from November 1^{st}, execution #2 had input of data two days, November 1^{st} and 2^{nd} together, and so on). The performance of the classifiers as a function of the monitoring time was measured, and modeled using the *Gompertz function*, a widely used function in the parametric form :

$$y(t) = ae^{be^{ct}}$$

(for different values of a, b and c for each attribute, according to the evolution of the learning curve of its classifier).

The applicability of the *Gompertz function* for modeling local behavior patterns of mobile users was demonstrated in [23], predicting the applications users chose to install. This experiment had shown that this process can be best modeled using the *Gompertz* instance for $a = 1$, $b = c = -1$. It has also been used for modeling mobile phone uptake [24], population in a confined space [25], and growth of tumors [26].

Following is a detailed description of 4 of the classifiers we have created: (1) the ethnicity of a user, (2) whether the user is a student or not, (3) whether the user is a native US citizen or not, and (4) who is the significant other of each user.

Ethnicity: The *Louvain* method for community detection [27] partitioned the SMS social network into 13 disjoint groups (Figures 1 and 2), successfully predicting the ethnicity of 60% of the participants (77 out of 128 with known ethnicity). See prediction evolution and *Gompertz* regression in Figure 3.

Is student: Our dataset contained the occupation of 88 users, almost half of which were students. *Rotation-Forest* classifier yielded AUC of 0.639 and an F-measure of 0.625. See prediction evolution and *Gompertz* regression in Figure 4.

US-natives: Our dataset contained information regarding the origin of 86 user. Our *Naive-Bayes* classifier yielded AUC of 0.728 and an F-measure of 0.806. See prediction evolution and *Gompertz* regression in Figure 5.

Significant other: Analyzing the social structure of the Bluetooth collocation social graph (Figure 6) succeeded in classifying 65.6% of the couples (44 out of 67). See prediction evolution and *Gompertz* regression in Figure 7.

V. Discussion and Conclusions

In this paper we have examined the way learning and prediction process evolves in time, as the amount of data available to the learning algorithm increases. We have shown that this process can be well modeled using the known Gompertz function. We have demonstrated that this result holds for the prediction of different features, both social and individual, and for a set of different prediction methodologies, using a varying number of input signals, all collected via mobile phones in a

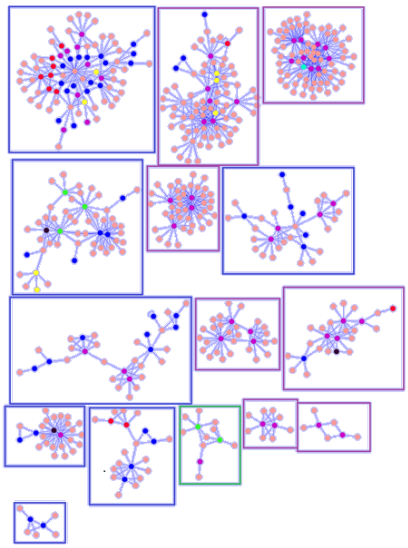

Fig. 1. Partitioned SMS Social Network Using *Louvain* algorithm [27]. Each group has different ethnicity according to the major ethnicity of the group (Blue: Asian, Purple: Caucasian, Green: Middle Eastern).

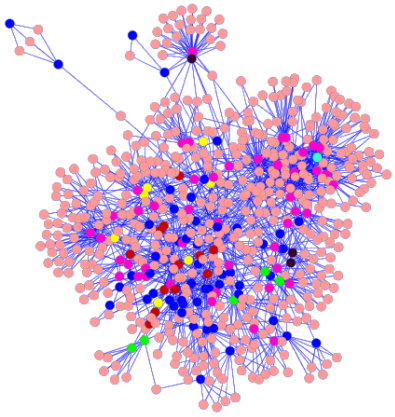

Fig. 2. SMS Social Network Graph created over 65 weeks (graph also includes unknown out-of-study nodes, which connect to at least two known in-study nodes). Different vertex colors represent different ethnicity.

field deployment. We have done so using a unique dataset, cultivated from a long-term comprehensive study done at the MIT dorms.

Furthermore, our findings can be used as a method for advance prediction of the maximal learning accuracy possible for the learning task at hand, using just the first few measurements, by extrapolating the learned Gompertz functions as illustrated in Figure 8. Such extrapolation can either be used in design-time to predict the maximal expected accuracy, or alternatively to assess in real-time our location for each signals estimated accuracy curve. We can then use this information to evaluate the analysis method, anticipate the timeline for increased accuracy, and understand when it is time to stop collecting data as we have reached a state of saturation. Another possible use is comparing different learning processes to one another, and using this information as part of the experiment or analysis

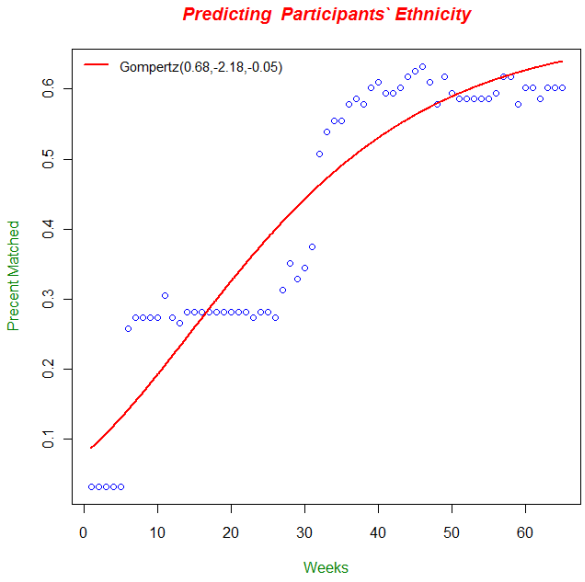

Fig. 3. The prediction accuracy of the *Ethnicity* classifier. The vertical axis represents the percentage of correct predictions, for the Gompertz function $f(t) = 0.68e^{-2.18e^{-0.05t}}$ with regression residual standard error of 0.06676, and achieved convergence tolerance of 5.568e-06.

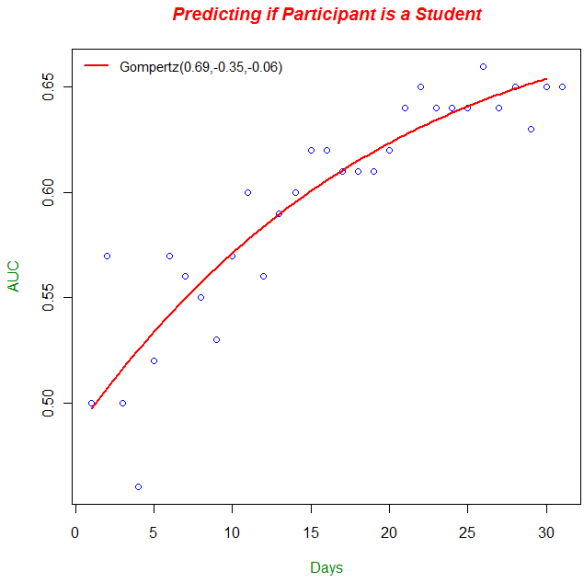

Fig. 4. The prediction accuracy of the *Is Student* classifier. Vertical axis represents AUC values. The fitted Gompertz function is $f(t) = 0.69e^{-0.35e^{-0.06t}}$ with regression residual standard error of 0.02237, and achieved convergence tolerance of 4.095e-06.

management process.

Correlations between the evolution trends of the different learning process, as depicted in Figure 9, may imply certain underlying correlations in the raw data itself, and can be used as additional validation for correlated features and observations (such as the suggestion that people might have

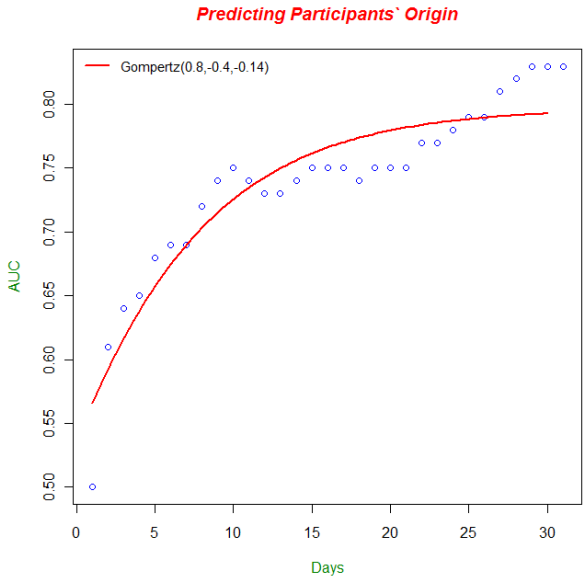

Fig. 5. The prediction accuracy of the *US Citizen* classifier. The vertical axis represents the area under curve (AUC) values. The best fitted Gompertz function is $f(t) = 0.8e^{-0.4e^{-0.14t}}$ with regression residual standard error of 0.02591, and achieved convergence tolerance of 7.404e-06.

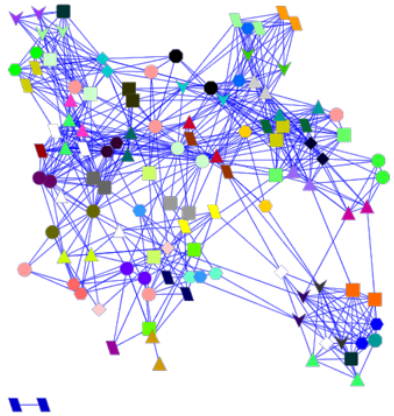

Fig. 6. Bluetooth social network graph of face-to-face interaction during November 2010. Significant others have the same shape and color. Each link represents at least 100 interaction.

a higher tendency to marry within their own ethnic group, as has been widely observed [28], [29]. In addition, this information could be used for informing the design of data collection configurations for ongoing or future data collection initiatives. For example, if two features are highly correlated, yet one of them is much "cheaper" to extract (e.g. requires only reading the phones built-in call-log database, compared to battery-intensive GPS scanning), it might be decided that extracting the cheaper feature alone is sufficient, deducing the more expensive feature using their correlation. Alternatively, we might want to actually make sure that two correlated values are gathered in order to strengthen the result and help deal with noise.

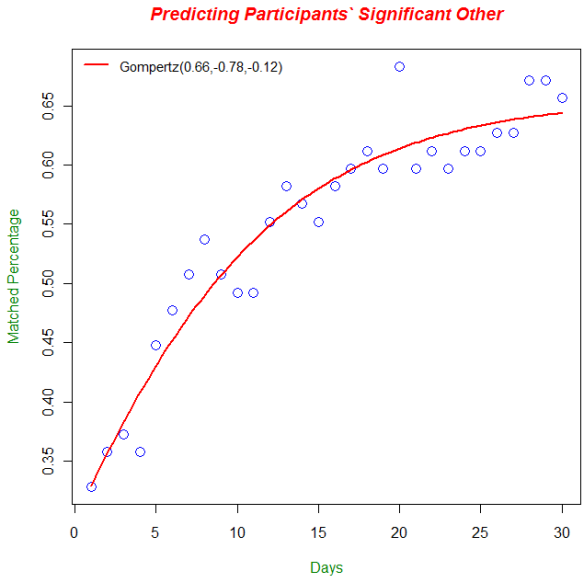

Fig. 7. The prediction accuracy of the *Significant Other* classifier. The vertical axis represents the percentage of correct matches. The fitted Gompertz function is $f(t) = 0.66e^{-0.78e^{-0.12t}}$ with regression residual standard error of 0.02762, and achieved convergence tolerance of 1.505e-06.

At this point it is interesting to mention the work of Dey et al. [30] that have shown that people statistically carry their phones with them much less than they might think. This might explain why there are saturation limits in learning accuracy of mobile phone data, as Bluetooth proximity based analyses assume that the phone is an accurate proxy for its owner and is located where the owner is.

It should be noted that our main goal in this study was to investigate the learning process over time, rather than evaluate the specific models and how they generalize. In future work we intent to return to each of these models, evaluating it in details. We are also continuing our investigation of the properties of learning and prediction of human and social constructs based on mobile phone gathered data.

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

References

[1] N. Eagle, A. Pentland, and D. Lazer, "Inferring social network structure using mobile phone data," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106, pp. 15 274–15 278, 2009.

[2] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, 2011.

[3] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne, "Social science: Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[4] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

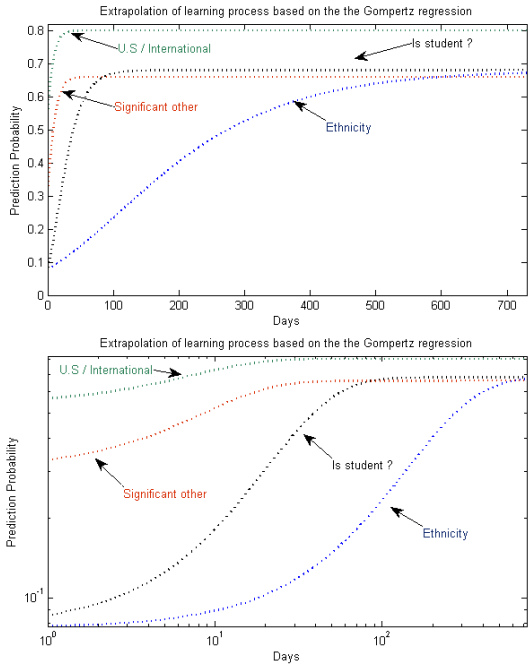

Fig. 8. Extrapolation of the learning process based on the Gompertz regression for the four learning tasks, in linear scale (top) and log-log scale (bottom).

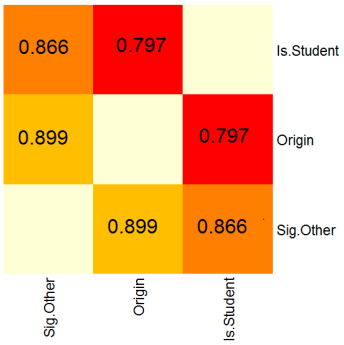

Fig. 9. Pearson correlation between the learning process dynamics of three traits. Although the evolution trajectories of the learning processes are positively correlated, while some are very highly correlated (e.g. Origin vs. Significant other), which might point out a strong correlation in the underlying data itself (i.e. people tend to get married more within the same ethnic group), other display lower correlation (e.g. Origin vs. Is student).

[5] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[6] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[7] C. man Au Yeung, M. Noll, C. Meinel, N. Gibbins, and N. Shadbolt, "Measuring expertise in online communities," *Intelligent Systems, IEEE*, vol. 26, no. 1, pp. 26 –32, jan.-feb. 2011.

[8] N. Eagle and A. Pentland, *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268.

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 06 2008. [Online]. Available: http://dx.doi.org/10.1038/nature06958

[10] N. Eagle, M. Macy, and R. Claxton, "Network diversity and economic development," *Science*, vol. 328, no. 5981, pp. 1029–1031, 2010.

[11] A. Madan, M. Cebrian, D. Lazer, and A. Pentland, "Social sensing for epidemiological behavior change," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, ser. Ubicomp '10. New York, NY, USA: ACM, 2010, pp. 291–300. [Online]. Available: http://doi.acm.org/10.1145/1864349.1864394

[12] "Sense networks. http://www.sensenetworks.com/."

[13] A. Madan, K. Farrahi, D. G. Perez, and A. Pentland, "Pervasive sensing to model political opinions in face-to-face networks." Springer, 2011, pp. 214–231.

[14] R. Montoliu and D. Gatica-Perez, "Discovering human places of interest from multimodal mobile phone data," in *Proc of 9th Int. Conference on on Mobile and Ubiquitous Multimedia (MUM,',',')*, 12 2010.

[15] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The jigsaw continuous sensing engine for mobile phone applications," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '10, New York, NY, USA, 2010. [Online]. Available: http://doi.acm.org/10.1145/1869983.1869992

[16] A. Joki, J. A. Burke, and D. Estrin, "Campaignr: A framework for participatory data collection on mobile phones," 2007. [Online]. Available: http://www.escholarship.org/uc/item/8v01m8wj

[17] T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich, "Mobiscopes for human spaces," *IEEE Pervasive Computing*, vol. 6, pp. 20–29, 2007.

[18] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 39, no. 1, pp. 43–55, 2009.

[19] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[20] A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 251–260.

[21] L. Rokach, M. Kalech, I. Blank, and R. Stern, "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.

[22] e. a. Hall, M., "The weka data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[23] W. Pan, N. Aharony, and A. Pentland, "Composite social network for predicting mobile apps installation," in *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)*, 2011, pp. 821 – 827.

[24] P. Rouvinen, "Diffusion of digital mobile telephony: Are developing countries different?" *Telecommunications Policy*, vol. 30, no. 1, pp. 46 – 63, 2006.

[25] G. Erickson, P. Currie, B. Inouye, and A. Winn, "Tyrannosaur life tables: An example of nonavian dinosaur population biology," *Science*, vol. 313, no. 5784, pp. 213–217, 2006.

[26] A. d'Onofrio, "A general framework for modeling tumor-immune system competition and immunotherapy: Mathematical analysis and biomedical inferences," *Physica D*, vol. 208, pp. 220–235, 2005.

[27] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, 2008.

[28] M. Kalmijn, "Intermarriage and homogamy: Causes, patterns, trends," *Annual review of sociology*, pp. 395–421, 1998.

[29] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.

[30] A. Dey, K. Wac, D. Ferreira, K. Tassini, J. Hong, and J. Ramos, "Getting closer: an empirical investigation of the proximity of user to their smart phones," in *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 2011, pp. 163–172.