Show Me Your App Usage and I Will Tell Who Your Close Friends Are: Predicting User's Context from Simple Cellphone Activity

Alain Shema
School of Information Studies
Syracuse University, Syracuse, USA
sralain@syr.edu

Daniel E. Acuna
School of Information Studies
Syracuse University, Syracuse, USA
deacuna@syr.edu

Abstract

Personal interactions using cell phones are so embedded in our everyday lives that they go almost unnoticed. We may try to protect ourselves from releasing sensitive information by increasing privacy protections, but how much can be inferred from our most basic phone usage? Using a large-scale annotated dataset of cell phone usage, we build a predictor to determine location context (home, work, commute) and social relationships (with close friend, with family) based on the clock of the phone and sequences of apps executed. Surprisingly, we show that just using this basic information we can accurately predict whether someone is at home, at work, and/or with close friends, family. We note that this is almost inevitable because it only depends on *using* the phone and not the privacy settings. Our results suggest that our relationship with technology gives away more than we might suspect. This presents opportunities and challenges discussed in this paper.

Author Keywords

Sequential decision making; Cyber-human systems; Cell phone usage

ACM Classification Keywords

C.1.3 [Other Architecture Styles]: Cellular architecture (e.g., mobile); K.4.1 [Public Policy Issues]: Privacy

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3053275

Introduction

Increasingly, technologies are essential for achieving our daily goals. While this opens many possibilities, it also makes us vulnerable to tracking. For example, researchers have shown that anonymous transaction networks can be statistically de-anonymized using few leaks [8, 10, 4]. As a consequence, consumer advocacy groups have pushed for more stringent restrictions on many technologies, such as cell phones, to disable such tracking. However, it remains unclear how much information we give away even when sensors and GPSs are disabled. In this work, we ask how much we can infer about a user based on large-scale regularities in basic usages of cell phones. Given that cell phones are one of the most ubiquitous technologies in use today, this research opens up questions about how technology affects our daily lives.

Cell phones are a good example of how technology has become pervasive and how privacy concerns around them have evolved. Worldwide, cell phone subscriptions increased from 12% in 2000 to 98.6% in 2015 [1]. With their rich sensors, mobile phones enable the development of applications that consider user's context to provide a customized experience. For example, GPS sensors enable the use of maps that provide real-time navigation services to users. Previous research has proposed to use cell phones for activity recognition (e.g., [9]). However, it remains difficult to infer higher level contexts of user's location (work, restaurant, etc.); activities (sleeping, cooking, etc.); surroundings (with friends, with family, etc.). A few approaches have been suggested to infer user's context. For example, Gellersen et al. proposed the use of multiple generic sensors, especially location and audio, to infer user's context [5]. This approach, while effective, introduces a number of privacy concerns. For example, users do not like when their cellphone is constantly "listening" to them. There are increasing privacy concerns around the type of information that a cell phone can collect from users, and many regulations now prevent the scope of tracking [3]. Cell phones, therefore, are a good representation of pervasive and personal technology.

In this study, we investigate how to predict location context (e.g., home, work) and social relationships (e.g., family, friend) only by analyzing the clock and the sequence of applications being executed. This effectively bypasses privacy settings. We use the Mobile Data Challenge (MDC) dataset (see materials) and analyze close to 4 billion seconds of labeled phone activity. Surprisingly, we show that by using simple time-based features we are able to infer places and social relationship accurately. We discuss the implications of these findings and propose new avenues of research for the future.

Methodology

Data

In this study, we used the data collected by the Idiap Research Institute and the Nokia Research Center in Lausanne during their "Lausanne Data Collection Campaign (LDCC)" from 2009 to 2011. The data comes from close to 200 participants and was collected through phones equipped with a special software. The software captured data related to social interactions (call logs, text messaging, Bluetooth interaction and audio samples); user location (GPS, cellular network and WiFi); media creation and usage (the camera and music player); and other behavioral patterns (application usage and accelerometer sensor) [7]. The dataset contains more than 8 million application usage logs: time, type of events (application started, viewed, closed, etc.), name of the application, and others. There are records about the places the user visited, including the type of location, the time they arrived and left the place; whether they were in company of friends, family, colleagues, etc. The dataset specifies 10 different types of location: (1) home, (2) home of a friend, (3) workplace/school, (4) location related to transportation, (5) workplace/school of a friend, (6) place for outdoor sports, (7) place for indoor sports, (8) restaurant or bar, (9) shop or shopping center and (10) holiday resort. With the number of apps continuously increasing, we grouped the apps by function (e.g., SMS and Skype are "communication" apps) and we use such function to do the predictions.

Method

We used Random Forest to perform the prediction. Random forest is a popular statistical method that combines several decision trees on different subsets of the data and features. This combination increases the accuracy of the predictions and reduces the variance of the estimators (see [6]).

We developed two models. The first model (baseline) uses as features the day of the week, the hour of the day and the week of the year. Intuitively, people go to certain locations based on time. For example, people generally go to work during week days rather than weekends. Thus, time is a good indicator for someone who would want to randomly guess a person's location. For the second model, we use the features of the baseline model, and also the five previously executed applications and the durations of their executions.

To train and validate the models, we split the users into 60% and 40%, respectively. The split is done such that the users in the training set do not appear in the validation set, and vice versa. Thus, the trained models are tested on users that they have not seen before. Though this approach potentially reduces the accuracy rates obtained, it makes the models more generalizable since they can be applied to previously unseen users.

Results

The first step to understand statistical relationships between variables is to explore base rates occurrences. The dataset has ten types of location contexts but the distribution of time among them is very skewed with the top two location contexts gathering more than 90% of the usage time. More concretely, users spent about 60% of their time at home, 32% at work, and the rest doing leisure activities (see table 1). Additionally, the dataset has recorded four types of social relationships. Users spent 26% of their time with close friends, 49% with friends, 33% with colleagues, and 9% with incidentals. The difference between location contexts and social relationships is that the latter can happen simultaneously — e.g., being with family and friends at the same time. Based on these numbers, a model that does not use any features can expect to predict the location context with an accuracy of 60% given that people are most of the time at home. Similarly, we expect to predict that someone is with their family 50% of the time, and so on. Therefore, these base rates form a first step to understand basic performance for the models we present below.

One of the goals of this research is to show how much we can predict the location context of a user based on simple app usage data. To test this hypothesis, we propose a model that uses features related to date and time, as well as features related to previous apps being used. The time features in particular are the day of the week (Monday = 0 through Sunday = 6), hour of the day (0 to 23), and week of the year (0 to 51). These features are meant to capture some intuitive trends, e.g., people are more likely to be at home during night time and at their workplace on Monday mornings. We split the data with 60% of the users in the training and 40% for validation. Using a Random Forest as classifier with 500 trees (see methods), the model predicts location with 65% accuracy, which is significantly bet-

Location context	%
Home	60.19
My workplace/school	31.69
Home of a friend	6.18
Place for outdoor sports	0.52
Place for indoor sorts	0.39
Workplace/school of a friend	0.31
Restaurant or bar	0.27
Shop or shopping center	0.19
Holiday resort	0.12
Location related to transportation	0.09

Table 1: Distribution of location across users. Perhaps, not surprisingly, most users spent their time at home and at the workplace, accounting for nearly 92% of the time recorded. This makes the prediction of a simple model to have around 60% accuracy without looking at features

ter than a model that does not use any features — paired t-test $t(30) = 2.02, p = 0.047$ (Fig. 1). Looking at feature importance from Random Forest [6], we find that the most important feature in the prediction is the hour of the day, followed by week of year (Fig. 2). This shows that even with simple features like this, we can predict better location contexts relatively accurately.

We additionally test whether adding features related to apps and their usage would increase accuracy. For example, during the weekends most users are at home but repeated usage of email and calendar apps may reveal that the user is at work. Thus, this information should increase the accuracy of the prediction. The model, therefore, incorporates as features the last five apps executed by the user and the times those apps were used. Again, we use a Random Forest classifier with 500 trees to predict the location context.

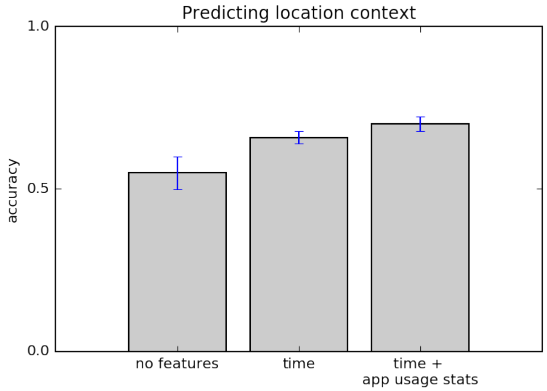

Figure 1: Model performances for predicting location context. No features just uses the base rates, time uses time of day, day in week, and week in year. App usage stats uses the last five apps viewed and the usage timings.

This mode achieves 69% accuracy (Fig. 1). This suggest that information about the sequences of usage improves our ability to predict user behavior.

Finally, we test whether our models can predict the social relationships of the user. For example, depending on the time of the year and hour of the day, we can statistically determine whether the user is with friends or family. Additionally, if we know that a user has not executed an application for a long time, then that might indicate that the user is with a close friend and he or she does not want to spoil an entertaining conversation. Our models are effective at predicting social relationships to an extend. In particular, a model that uses only time would predict that a user is with a close friend, family, colleague, or incidental with 73%, 60%, 70%, and 93% accuracy, respectively. The model which uses the apps and their usages in addition to time achieves

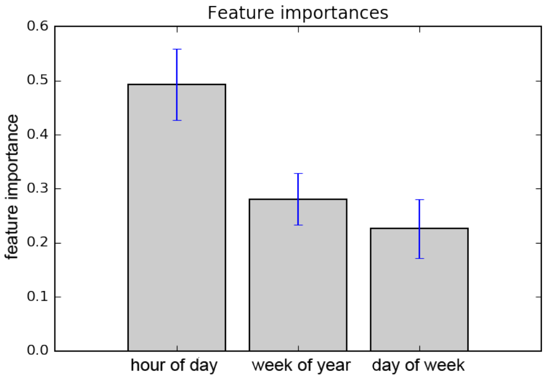

Figure 2: Simple model feature importance

even higher prediction accuracies of 82%, 65%, 73%, and 95% (Fig. 3). With no features, we would have expected to make this predictions at 74%, 51%, 67%, and 91% rates. This suggests that the way we use the phone gives good clues to understand our relationship to someone.

Discussion and Conclusion

In this research, we set out to explore how much we can infer about users activities from cell phone app usage. By using a large scale dataset of mobile usage activity, we built a statistical model to predict the type of location of a user (e.g., home, park, work) from the date and time of the app usage, the usage sequence of apps, and the time of usage for each app. This model is 18% more accurate (from 58% to 68% accuracy) than a model that does not use any features. We were also able to predict the relationship of the user in a social context. While users of cell phones have increasingly more privacy options, our approach does not use any monitoring tools or sensors. As technology becomes more pervasive, it is important to explore these implications.

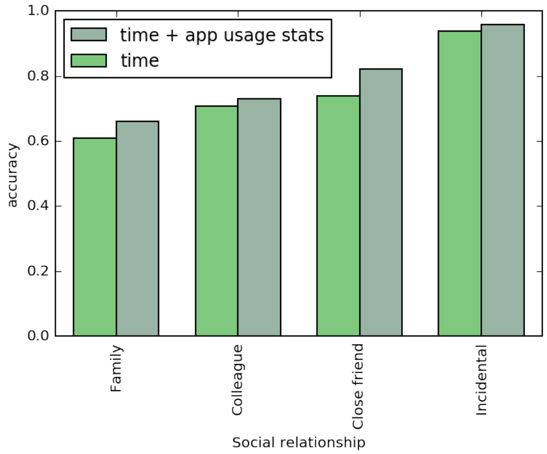

Figure 3: Comparison of models for predicting social relationships. Even though all base rates of these activities occur less than 50% of the time, the models are much more accurate.

Our approach is able to infer the location context and human surrounding of our users at any point in time. The variability across users hindered our statistical model to the point where the differences between models was not statistically significant. However, there are many avenues for improving the across-user performance. For example, our model did not use any information about the users on whom we made predictions. If we are able to see small piece of information about a user (e.g., asking once whether the current place is home), we should be able to boost performance much more. While our approach does infer location relatively successfully, future work will explore how to incorporate user-specific feedback to make this even more accurate and find out which specific piece of information increases accuracy.

Our analysis and statistical model relied on a dataset that was created several years ago and in a geographical location that might not be representative. For example, the number of applications available for iPhone on the Apple App Store has exploded 3-fold between June 2012 and June 2016 [2]. Also, the users in the MDC dataset all come from Lausanne, Switzerland. This introduces questions about the external validity of the results as there might be significant behavioral changes from the users in this dataset and users from other places (e.g., US). We tried to control for this disparity by grouping the apps into categories (e.g., texting, social apps, etc.). Nowadays, technology is even more pervasive and there is an app for almost anything. Thus, we suspect that this should make our approach even more successful today. In future work, we plan to explore the effectiveness of our methods on other datasets from other geographical locations.

Overall, our approach reveals that we can learn much about users using only the bare minimum logs of apps usage. Other research has shown that we can infer users activity using monitoring tools that are significantly more invasive or from information about sets of users. Our approach relies on data whose collection is almost impossible to restrict. These results open a host of questions about the limitations of privacy settings and approaches, and introduce new ways of customizing user's experiences. Thus, learning users' context from simple usage activity is a technique that we think will become more common in the future as people restrict data collection by sensors.

Acknowledgements

This study was conducted using the Mobile Data Challenge (MDC) dataset. We would like to extend our gratitude to the team at the Nokia Research Center in Lausanne and Idiap. This research was supported in part through computational resources provided by Syracuse University, supported by the NSF award ACI-1541396. Daniel E. Acuna was partially supported the National Science Foundation under Grant No. 1646763. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the supporting organizations.

References

[1] 2017. Mobile cellular subscriptions (per 100 people). (2017). http://data.worldbank.org/indicator/IT.CEL.SETS.P2

[2] Sam Costello. 2017. How Many Apps Are in the App Store? (Oct 2017). https://www.lifewire.com/how-many-apps-in-app-store-2000252

[3] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013), 1376.

[4] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, and others. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347, 6221 (2015), 536–539.

[5] Hans W Gellersen, Albercht Schmidt, and Michael Beigl. 2002. Multi-sensor context-awareness in mobile devices and smart artifacts. *Mobile Networks and Applications* 7, 5 (2002), 341–351.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer New York Inc., New York, NY, USA.

[7] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin* (2010).

[8] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*.

IEEE, 111–125.

[9] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. 2005. Activity recognition from accelerometer data. In *AAAI*, Vol. 5. 1541–1546.

[10] Mudhakar Srivatsa and Mike Hicks. 2012. Deanonymizing mobility traces: Using social network as a side-channel. In *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 628–637.