Poster: A Study of Smartphone User Privacy from the Advertiser's Perspective

Yan Wang[†], Yingying Chen[†], Fan Ye[‡], Jie Yang[§], Hongbo Liu[♯]
[†] Department of ECE, Stevens Institute of Technology, Hoboken, NJ, USA
[‡]Peking University, Beijing, China; [§]Department of CSE, Oakland University, Rochester, MI, USA
[♯]Department of CIT, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA
[†]{*ywang48, yingying.chen*}*@stevens.edu*; [‡]*yefan@pku.edu.cn*;[§]*yang@oakland.edu*; [♯]*hl45@iupui.edu*

I. Introduction

The huge success of smartphones is largely fueled by the availability of millions of phone apps that provide functions covering all aspects of our lives. A large portion of these apps are free. Their developers get financial support from advertisers by embedding their advertisement libraries to display mobile advertisements to users. To gain better understanding of user habits and behaviors for accurate ad targeting, these apps customarily scavenge private user data, ranging from the phone's IMEI number, MAC addresses of nearby access points, the user's location, even the contact list, and send it to advertisers [1]. Ultimately, these "free" apps are not entirely free: users pay the price of their privacy.

In this work, we seek to answer an important question: how much does the advertiser know about the user, in particular, her social and community relationship (e.g.,family, colleagues and friends) from the leaked private data? We focus on one important aspect of that perspective, the *social relationship* of a user, which is defined as a pair-wise relationship between two users with certain kind of physical interactions such as colleagues, families, and friends (not virtual friends from online social networks). In particular, we quantify to what extent an advertiser can learn and infer users' relationships by developing a privacy leakage inference framework. To facilitate the understanding on the consequences of privacy leakages, we take a two-step approach: privacy leakage inference and profile modeling via experimental study, and inference framework evaluation via trace-driven study, as depicted in Figure 1.

Our systematic study on privacy leakage inference involves both real experiments with multiple volunteers as well as trace-driven studies with human mobility traces obtained from the Foursquare trace [2]. We verify that by using 3 weeks of private data from the trace, an advertiser can infer colleague-based relationships of regular users at around 90% accuracy.

II. User Privacy Leakage Modeling And Experimental Study

A. Privacy Leakage Modeling

To understand the consequences of privacy leakages when an advertiser combines the data received from multiple apps and different users, we develop a three-layer privacy leakage inference framework (as depicted in Figure 1) including *Privacy Leakage Aggregation*, *User Connection Derivation* and *Relation Inference*. In addition, we introduce an important concept *connection*, which exists when two users share similarities in leaked data. The *connection* helps bridge the gap between raw privacy leakage data and higher level relationship inference. The intuition is that each type of particular relationship has certain temporal-spatial patterns in users' physical interactions, which can be captured by *connection*. For example, two family members usually stay together at home during late night and early morning; while classmates encounter each other frequently in classrooms during the daytime of weekdays.

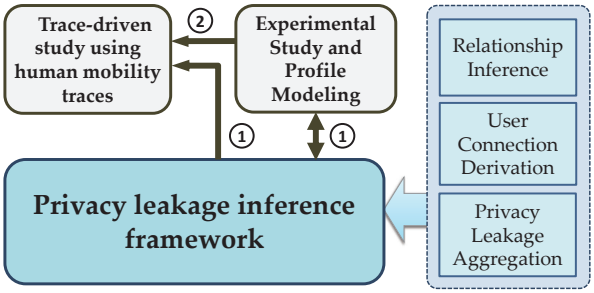


Fig. 1. We take a three-step approach to understand the advertiser's perspective on users' social and community relationships.

B. Experimental Study

Methodology. We conduct experimental study with 10 participants and their families for over one-month time period, among which five types of relationships exist: colleague, collaborator, classmate, friend, and family. We developed a tool to capture the privacy leakage information in real-time leveraging TaintDroid. During the experiments, we distribute smartphones to volunteers with the tool and the top 10 popular apps from 19 categories installed, and the volunteers are encouraged to use whichever apps they are interested in without knowing the purpose of this experiment. We find that an advertiser can infer over 90% of the pairwise relationships by utilizing a threshold-based approach based on *connections* (e.g., IMEI and GPS location).

C. Deriving Privacy Leakage User Profiles

Based on the experimental data collected over one month, we next build privacy leakage user profiles, namely *activeness based profiles*, to statistically capture the temporal and spatial patterns of privacy leakages of different users. The

privacy leakage user profiles quantify the users' privacy leakage characteristics, i.e., the leakage probability determines whether type i privacy leakage happens or not in a time window. The profiles will be used in our trace-drive studies in the next Section to facilitate the understanding of the user relationship inference from the advertiser's point of view.

Activeness Based Profile. The activeness based profiles are generated based on privacy leakages from each participant in our experiments. Assume there are N types of privacy leakages observed in total. We divide the time in day d into T time windows as $\{w_t, t = 1, \cdots, T\}$. Then within a time window w_t, a vector $\Phi^{u,d,t}$ is defined to capture the numbers of occurrences of different privacy leakage types collected from the user u, and each element $\Phi^{u,d,t}(i)(i = 1, \cdots, N)$ corresponds to the number of times privacy leakage type i occurs. For example, when $\Phi^{u,d,t}$ equals to $[2, 1, 0]$, it means 2 occurrences of leakage type 1, 1 occurrence of leakage type 2 and 0 occurrence of leakage type 3 in time window w_t at day d for user u.

We then define $\gamma_i^{u,d,t}$ to indicate whether the privacy leakage type i appears in the vector $\Phi^{u,d,t}$ or not, where $\gamma_i^{u,d,t} = 1$ when $\Phi^{u,d,t}(i) \neq 0$, otherwise $\gamma_i^{u,d,t} = 0$. The probability that type i leakage happens for user u in time window w_t across D days (e.g., $D = 7$ days) is defined as:

$$Prob_i^{u,t} = \frac{\sum_{d=1}^{D} \gamma_i^{u,d,t}}{D}. \quad (1)$$

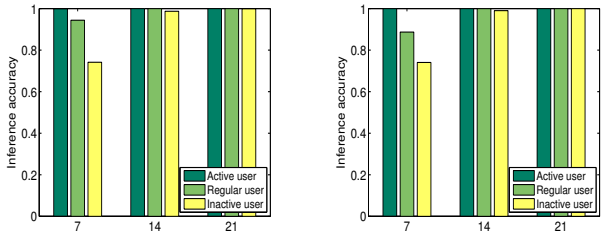
The activeness based profile of user u consists of $Prob_i^{u,t}$.

Categorization. Once the activeness based user profile is obtained, the advertiser could further categorize the profiles by the number of hours k_u the user u has privacy leakages in a one-day duration. We utilize two thresholding hours ρ_1 and ρ_2 with $\rho_1 > \rho_2$ to determine three representative user categories, namely *active user category*, *regular user category*, and *inactive user category*, as:

$$\alpha = \begin{cases} 1 \text{ (active user category), if } k_u \geqslant \rho_1; \\ 2 \text{ (regular user category), if } \rho_2 \leqslant k_u < \rho_1; \\ 3 \text{ (inactive user category), if } k_u < \rho_2. \end{cases} \quad (2)$$

III. Social Relationship Inference Leveraging Privacy Leakages

We systematically study the consequence of the privacy leakages obtained by advertisers by applying the privacy leakage model to a human mobility traces (i.e., the Foursquare trace [2]). In particular, we study how much an advertiser can infer about users' social and community relationships by utilizing the temporal and spatial patterns of connections to infer users' relationship. There are multiple privacy leakages that can produce connections between users. In this study we focus on the {user identity, location} combinations of most popular privacy leakages including IMEI, phone number, GPS location, Wi-Fi AP list, and network-based location. We distinguish two categories of relationships in the simulation as *Fact-based Relationship* and *Intelligence-based Relationship*, which covers traditional social relationships. The *Fact-based Relationship* includes colleagues, classmates, roommates, families that carry inherit similar, regular and repetitive spatial-temporal connection patterns as dictated by the relationship, whereas the *Intelligence-based Relationship* includes friends, which do not necessarily carry such patterns.



(a) Fact based relationship (b) Intelligence based relationship

Fig. 2. Inference accuracy, Foursquare mobility trace, (a) and (b) are from the activeness based profiles.

Figure 2 compares the accuracy of pairwise social relationship inference (for both fact and intelligence based relationships) by applying the activeness based profiles to the Foursquare trace under different sizes of observation windows. We use 3 days and 2 days as the threshold for fact-based and intelligence-based relationship inference respectively, which is introduced in Section II. From Figure 2 we observe that an advertiser can achieve over 80% inference accuracy for both fact-based and intelligence-based relationship. In addition, we observe that the inference accuracy decreases for less active users, which is reasonable since less usage leads to less privacy leakages. Furthermore, we find that longer observation windows help improve the inference accuracy, especially when the privacy leakage probability is low. It is because a longer window helps the advertiser to accumulate more data, resulting in more connections to identify users' relationships.

IV. Conclusion

This work serves as the first step towards a comprehensive understanding of the advertiser's perspective. In particular, we seek to discover what an advertiser can infer about users' social relationships by combining private data from many apps. We propose a privacy leakage inference framework that describes a general method for inferring users' social relationships. Our experimental study over one month demonstrates that an advertiser can infer 90% of users' social relationship correctly using simple heuristics. This observation is further confirmed by human mobility trace driven studies of a large scale data set. Concurrently, we are building a visualization tool (as a smartphone's App) that can capture and display the spatial-temporal statistics of privacy leakage to different advertisers in real time. We hope our work will eventually lead to a complete picture of the advertiser's perspective.

References

[1] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android permissions: User attention, comprehension, and behavior," in *Proceedings of the Eighth Symposium on Usable Privacy and Security*, pp. 1–14, 2012.

[2] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pp. 199–208, ACM, 2012.