Inferring Social Ties in Pervasive Networks: An On-Campus Comparative Study

Igor Bilogrevic
LCA1, EPFL
Lausanne, Switzerland
igor.bilogrevic@epfl.ch

Jean-Pierre Hubaux
LCA1, EPFL
Lausanne, Switzerland
jean-pierre.hubaux@epfl.ch

Kévin Huguenin
LCA1, EPFL
Lausanne, Switzerland
kevin.huguenin@epfl.ch

Philip Ginzboorg
Aalto University, Finland
Espoo, Finland
philip.ginzboorg@aalto.fi

Murtuza Jadliwala
Wichita State University
Wichita, KS, USA
jadliwala@cs.wichita.edu

Valtteri Niemi
Turku University, Finland
Turku, Finland
valtteri.niemi@utu.fi

Florent Lopez
Université Paul Sabatier/IRIT
Toulouse, France
florent.lopez@irit.fr

UbiComp'13 Adjunct, September 8–12, 2013, Zurich, Switzerland.
ACM 978-1-4503-2215-7/13/09.

http://dx.doi.org/10.1145/2494091.2494128

Abstract
WiFi base stations are increasingly deployed in both public spaces and private companies, and the increase in their density poses a significant threat to the privacy of users. Prior studies have shown that it is possible to infer the social ties between users from their (co-)location traces but they lack one important component: the comparison of the inference accuracy between an internal attacker (e.g., a curious application running on the device) and a realistic external eavesdropper (e.g., a network of sniffing stations) in the same field trial. We experimentally show that such an eavesdropper can infer the type of social ties between mobile users better than an internal attacker.

Author Keywords
Privacy; Social Networks; Mobile Networks; Inference

ACM Classification Keywords
C.2 [Computer-Communication Networks]: General; K.4.1 [Computers and Society]: Public Policy Issues—*Privacy*

Introduction
Upcoming device-to-device communication technologies (such as WiFi Direct, Nokia Instant Community (NIC) and Qualcomm's AllJoin) will enable opportunistic and localized interactions among physically co-located devices. Already today, Apple's AirDrop allows device-to-device

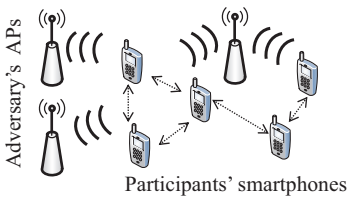

Figure 1: Trial network architecture.

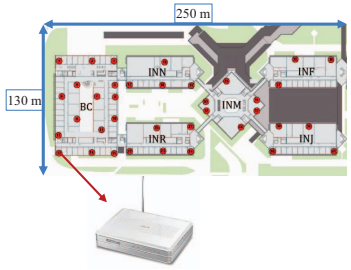

Figure 2: Map of the deployed wireless network of 37 APs controlled by the adversary.

exchange of data through the WiFi and there are numerous games that use Bluetooth for multiplayer.

The information that can be inferred only by observing the data exchange and users' (co-)location traces is very sensitive, in particular social ties [1]. Several studies have shown that (co-)location traces alone reveal the existence and type of social ties among people [2]. Most of the existing works are based on a single source of information: either proximity logs stored on the mobile devices or WiFi/cellular data obtained from the network operators.

We address the problem of social ties inference by conducting a comparative analysis in a peer-to-peer wireless network deployed on the EPFL university campus. We collected communication information from 80 participants carrying smartphones during a four-month experiment. We gathered data from both the smartphones and from the WiFi sniffing network, comprised of a set of access points (APs) passively intercepting communications between the participants. This *data-source diversity* reflects the information that can be gathered by (i) a curious application running on a device and (ii) a network infrastructure owner which seeks to infer the social ties between its employees or a network of compromised (community) hotspots. We use an inference framework to evaluate the accuracy of the inference of social ties. Our evaluation provides empirical evidence about the difference between the adversary's accuracy and that of a curious application.

Experimental Setup and Data

In addition to the network deployment and the field trial, the participants also answered survey questionnaires to construct, together with academic records, the ground truth of social ties.

Each participant was identified by a fixed random ID assigned at the beginning of the trial. The devices periodically sent beacon messages and ran background services that collected, at regular intervals, information including the list of neighbors' MAC addresses. Whenever a participant's device connected to the Internet, the logs were uploaded on a centralized database.

We implemented a practical adversary who monitored a fixed area by using a limited number of wireless sniffing stations. Specifically, the adversary had access to a deployed wireless mesh network of 37 WiFi APs (Asus WL-500gP running OpenWRT Linux) in a specific region of the campus (Figure). Each AP passively eavesdropped on the P2P packets generated by the participants' smartphones.Note that this external adversary did not have access to any information stored on the mobile devices, and that at no time did the participants connect to the Internet through the adversarial APs. Because application-layer encryption was sometimes used, we assume that the adversary did not have access to the content of the messages.

Summary. In our experimental data, we have: (i) the *internal data*, constituted by local logs collected by the mobile devices and containing encounter data (list of neighbors' MAC addresses, the timestamps and the RSSI values of received packets), and (ii) the *adversarial data*, containing the headers of the sniffed packets (sent by the mobile devices), which include the timestamps and RSSI values of received packets at the sniffing APs, as well as the device ID of the sender. We considered three classes of social ties: (i) *friends*, (ii) *classmates* and (iii) *others* (people can be simultaneously classmates and friends). To construct the ground truth data, each relationship is assigned one (or more) class labels based on (i) the

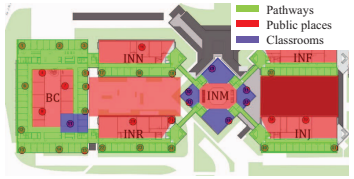

Figure 3: Partition of the map in *pathways*, *public spaces* and *classrooms*.

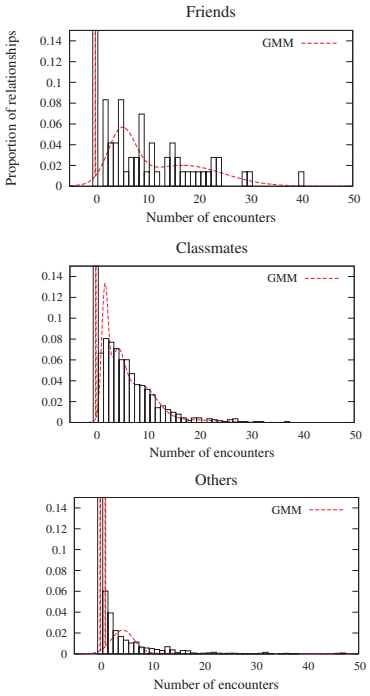

Figure 4: Experimental distributions and fitted pdfs from the adversarial data.

participants' answers to the survey questionnaires (for friends) and (ii) the database of academic course registration provided by the university (for classmates). We considered asymmetric social ties as symmetric because of the subjective nature of the friendship relationship that could be perceived by only one of the two persons. We observed that 25% of all the 3,160 possible relationships were tagged as *classmates* and 2% as *friends*. We tagged the remaining as *others*.

Social Ties Inference Framework

We developed an inference framework to evaluate the accuracy of the attacker's reconstruction of users' ties.

A relationship is defined by a set of encounters: we define an encounter between any two users as a continuous time interval that meets a threshold, both on its duration and on the proximity of the two involved users. The proximity between users is computed in two ways, depending on the data source. From the internal device logs, we use the RSSI values of messages exchanged between users in order to determine their distance, whereas from the adversarial sniffed data, the proximity is determined by computing the distance between the estimated location coordinates of each user, based on device-to-AP RSSI values.

A number of features are extracted from the internal/adversarial logs, at the granularity of an encounter or of a relationship, and fed to the classifier (Figure 5). Typical discriminating features include encounter duration, the proximity of users during encounters, inter-encounter time, and the number and periodicity of encounters [3].

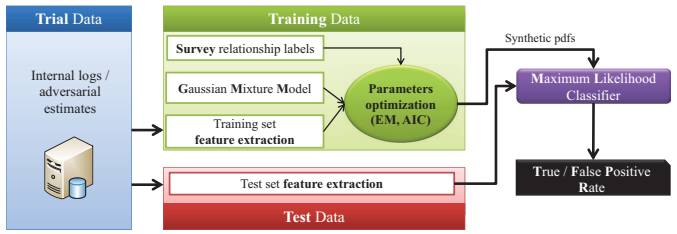

Figure 5: Overview of the inference framework.

In our context, the location where encounters take place is a key feature of relationships. This information is only available in the adversarial data and is computed from the users' locations estimated by the adversary. More specifically, we partitioned the region of the experiment covered by the adversarial network into three different areas (Figure 3).

Summary. At the granularity of an encounter, we extract: (i) the **encounter duration** (the time elapsed between the first and last message exchanged during the encounter), (ii) the **inter-encounter time** (the time elapsed since the end of the previous encounter and the beginning of the current encounter), and (iii) the **encounter RSSI** value (the average of the RSSI values of the messages exchanged during the encounter).

For the adversarial data, we further consider (iv) the **encounter location**, characterized by the fraction of the encounter time spent in each type of location (Figure 3). At the granularity of a relationship, we consider an additional feature: the total **number of encounters** over the trial. Figure 4 depicts the experimental histograms of one feature in our dataset.

Results

The performance of the classifier (Figure 6) is evaluated in terms of its *true positive rate* (TPR) and *false positive rate* (FPR).

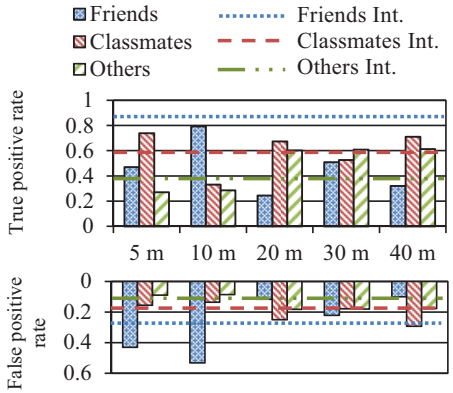

Figure 6: Accuracy of the external adversary in inferring social ties for different distance thresholds among devices (used in the encounter detection). The horizontal lines correspond to the internal data.

Using Internal Proximity Data. Using only 30% of data samples for training, we observe the best TPR (84%) and a moderate FPR (27%) for *friends*, which means that the actual proximity data and the encounter features we selected are quite discriminating for this class. The actual proximity data works well for the friends and classmates ties inference, whereas it has a more limited success for other types of ties. This is not surprising, as proximity information, without taking into account the actual location of the interaction, is intuitively tied to inferring close and periodic interactions.

Using Adversarial (Estimated) Proximity Data. At a lower encounter-distance threshold (5-10 meters), the adversary has a very high TPR rate for classmates and friends (79% and 74% respectively) while it suffers from an elevated FPR as well (53% for *friends*). The accuracy for *others* is relatively low at 28%, meaning that the prediction suffers from a significant number of false positives for *friends* and *classmates*. For larger thresholds (20-40 meters), the accuracy for *others* doubles and is consistently larger than for the *friends*.

Internal vs. Adversarial For high distance thresholds (20-40 meters), the success in inferring friendships from external data decreases by at least 50% compared to the case where internal data is used. Meanwhile, there is a two-fold increase in TPR for the *others* category. Overall, we observe that the increase of the encounter-distance threshold for the adversary is highly beneficial for the detection of non-friendships. By adjusting the threshold based on the target relationship class, the external adversary obtains an inference performance comparable to that of the internal adversary who uses the actual proximity information available on the devices.

References

[1] Crandall, D., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. Inferring Social Ties from Geographic Coincidences. *Proc. Nat. Ac. Sc. 107* (2010), 22436–22441.

[2] Eagle, N., Pentland, A., and Lazer, D. Inferring Friendship Network Structure by Using Mobile Phone Data. *Proc. Nat. Ac. Sc. 106* (2009), 15274–15278.

[3] Yoneki, E., Greenfield, D., and Crowcroft, J. Dynamics of Inter-Meeting Time in Human Contact Networks. In *Proc. Int. Conf. on Adv. in Soc. Net. Analysis and Mining* (2009).