Assessing Location Privacy in Mobile Communication Networks

Klaus Rechert[1], Konrad Meier[1], Benjamin Greschbach[2], Dennis Wehrle[1], and Dirk von Suchodoletz[1]

[1] Faculty of Engineering, Albert-Ludwigs University Freiburg i. B., Germany
[2] School of Computer Science and Communication, KTH - Royal Institute of Technology, Stockholm, Sweden

Abstract. In this paper we analyze a class of location disclosure in which location information from individuals is generated in an automated way, i.e. is observed by a ubiquitous infrastructure. Since such information is valuable for both scientific research and commercial use, location information might be passed on to third parties. Users are usually aware neither of the extent of the information disclosure (e.g. by carrying a mobile phone), nor how the collected data is used and by whom.

In order to assess the expected privacy risk in terms of the possible extent of exposure, we propose an adversary model and a privacy metric that allow an evaluation of the possible privacy loss by using mobile communication infrastructure. Furthermore, a case study on the privacy effects of using GSM infrastructure was conducted with the goal of analyzing the side effects of using a mobile handset. Based on these results requirements for a privacy-aware mobile handheld device were derived.

1 Introduction

Mobile communication systems as well as location-based services are now both well established and well accepted by users. The combination of location determination, powerful mobile devices (so-called Smartphones) and ubiquitous network communication options provide a lot of useful new applications but also bring new challenges to the users' privacy especially when it comes to location disclosure.

There are various occasions and motives for location disclosure. In general, one can classify location disclosure types into two different categories based on trust relationship between involved communication peers (cf. Fig. 1). A very common situation today is when users exchange their whereabouts with location-based service providers for tailored and context-sensitive information. Exchanging position information within groups through social network services (SNS) is also gaining in popularity. These services usually involve informed users who are aware that they are disclosing their location data. Service providers as well as social peers are considered as partially trusted, at least for the specific communication context, as communication is voluntary and communication peers are known.

X. Lai, J. Zhou, and H. Li (Eds.): ISC 2011, LNCS 7001, pp. 309–324, 2011.

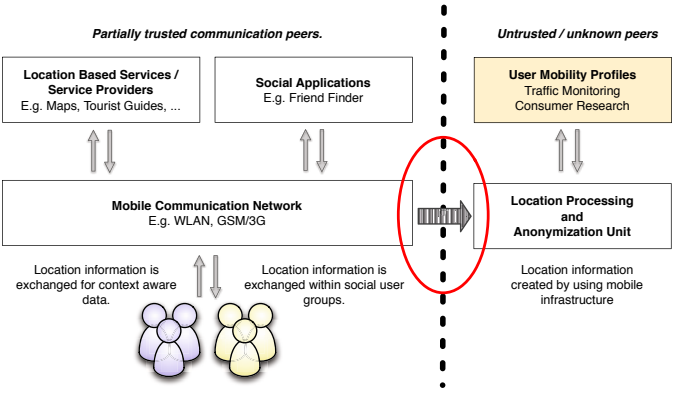

Fig. 1. Transfer of location information from (partially-)trusted peers to untrusted peers

Nowadays there is typically mobile communication involved. The communication infrastructure is usually also considered as partly trusted, i.e., there is a service agreement between the user and provider. Due to the specific nature of mobile communication networks, the location of mobile subscribers is known to the underlying infrastructure.

Hence, there is a second class of location disclosure, where location information from individuals is generated automatically or is observed by the infrastructure. Such information is valuable for scientific research [1] but also for commercial use (e.g. traffic monitoring[1] or location-aware advertising [2]). Therefore, location information might be passed on to third parties in an anonymized and / or aggregated way. In this case users are usually aware neither of the extent of their information disclosure (e.g. by carrying a mobile phone), nor how the collected data is used and by whom.

A lot of research has been done (e.g. [3]) on protecting a user's anonymity. However, when using mobile infrastructure users face two difficulties. First, due to regulations, the quality of service, but also technical conditions, location privacy protection measures like anonymity and obfuscation techniques seem inadequate or difficult to employ. Second, users suffer from limited and asymmetric knowledge. They have no knowledge on the nature, accuracy and amount of location data they have generated by using mobile communication infrastructure so far. As well, they cannot make a judgment about the level and quality of anonymization if the location data is exploited for various services. Since the data might be de-anonymized (e.g. [4]) or, even worse, if raw data leaks for whatever reason, users bear a latent privacy risk.

Thus, disclosing location data always conflicts with the user's privacy, since position information or movement history might lead to the user's identity. Col-

[1] For instance A1 Traffic (`http://www.a1.net/business/a1traffictechnologie` [12/1/2010]) or Vodafone HD Traffic (`http://www.vodafone.de/business/firmenkunden/verkehrsinfo-hd-traffic.html` [5/1/2011]).

lected location data can become a quasi-identifier, similar to a fingerprint [5]. Hence, by using external knowledge the identity of a specific user can be determined (e.g. [6]). Furthermore, through observing and evaluating a user's movements, his preferences and other possible sensitive information might be revealed. Such sensitive location-related data contains places a user visits frequently or at certain times and thus has special interest in. With the location data of other individuals his social relations become visible.

In order to assess the expected privacy risk in terms of the possible extent of exposure, first an adversary model has to be defined. Based on this model a simple privacy metric is proposed in order to assess the privacy loss in mobile communication networks and provide the groundwork for developing requirements for a privacy-aware communication device. As an example, we analyzed the impact on the user's privacy of different network configurations found at four GSM telephony infrastructure providers.

2 Related Work

Recent analysis of mobile phone call data records (CDR) showed that even sporadic anonymous location data with coarse spatial resolution contain sensitive information and could lead to possible identification. Humans tend to move in very regular patterns. A study using six months of call data records showed that humans stay more than 40% of the time at the same two places [1,7]. In another study on anonymized aggregated call data records, the movement patterns of commuters in two cities were compared [8]. Similar studies were conducted on tourist movement patterns in New York and Rome [9,10].

Sohn et al. analyzed GSM data to determine a user's movement mode based on radio signal fingerprints. The authors were able to distinguish between walking, driving and stationary profiles with a success rate of about 85% [11]. De Mulder et al. conducted a study on the possibilities of re-identification of individual mobile phone subscribers based on available cell data. In their study they evaluated a Markovian model and a model based on the sequence of cell-IDs. They report a success rate of about 80 % for the latter method [12].

From the aforementioned studies one could conclude that using a mobile communication network (e.g. GSM) is a threat to a user's privacy. However, from a user's perspective the question remains how much knowledge a network provider has on his movement patterns and which of the available networks pose the least threat for his privacy. Lee et al. dealt with location privacy in GSM networks only on the protocol layer in the relation between mobile station, VLR and HLR. However, location data as a possible quasi-identifier was not discussed [13]. Ardagna et al. introduce a scenario with a semi-trusted (mobile) network provider and propose a multi-path communication approach to achieve k-anonymity for the initiating sender of a message [14]. Their approach relies on a hybrid network infrastructure where subscribers are able to form ad-hoc networks. However, such an approach protects only the relation between sender and final recipient (e.g. LBS-provider).

In order to evaluate a privacy metric an adversary model is required. A popular model is an adversary that observes in some way generalized location data and tries to reconstruct this data to connected traces of a single individual. In a second step the adversary may re-identify the traced individual through his workplace or home by incorporating external knowledge (e.g. [15]) For instance, Shorki et al. defined a location privacy metric that measures the (in)ability of an adversary to accurately track a mobile user over space and time [16].

A different method for measuring location privacy is to make use of the uncertainty of an adversary in order to assign a new observation to a trace of a specific individual, e.g. by assigning probabilities to movement patterns and thus compensating changed pseudonyms [17]. A similar measurement was proposed as *time-to-confusion* metric, the tracking time of an individual until the adversary cannot determine the next position with sufficient certainty [15]. The (un)certainty measure is based on the entropy of the observed position and the probability of the expected or calculated next location. Sending a variety of locations for each query also increases the ambiguity and thus the level of privacy [18].

However, in most mobile communication relations the anonymity assumption seems inadequate. Furthermore, the aforementioned privacy metrics usually require full insight into the set of all users to determine the level of privacy for a single user within this set.

3 Location Privacy in Mobile Telephony Networks

When using mobile communication, location data is generated, accumulated and stored, as a technical and possibly legal[2] necessity. As an example we discuss the GSM infrastructure, because it is widely deployed and recently software and analysis tools have become available [3,4]. Its successors UMTS (3G) and LTE (4G) still share most of its principal characteristics. The goal is to uncover hidden privacy risks posed by the network's background communication and the effects of different network configurations on the user's privacy. In contrast to active usage (e.g. phone calls, texting), location data is not always provided voluntarily.

A typical GSM network is structured into cells, served by a single base station (BTS) and larger cell-compounds called location area (LA). In idle state no dedicated channel is assigned to the mobile station (MS). It only listens to the common control channel (CCCH) and to the broadcast control channel (BCCH) [19,20] and is otherwise in standby mode to save energy. Through System Information Messages on the BCCH the MS receives periodically a list of neighboring cells from the serving BTS and performs signal strength measurements on these base stations. This way the MS can always select the BTS with good received

[2] For instance EU Directive 2006/24/EC (Data retention), `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0024:en:NOT`, [5/21/2011]

[3] Open Source GSM Baseband implementation, `http://bb.osmocom.org` [05/19/2011]

[4] Open Base Station Controller OpenBSC, `http://openbsc.osmocom.org` [05/19/2011]

signal strength in order to maintain network attachment. To establish a connection to the MS in case of an incoming connection, the network has to know if the MS is still connected to the network and in which LA the MS is currently located. To accomplish that, the location update procedure was introduced. Either periodically or when changing the LA, a location update (LU) procedure is performed. Within this procedure the phone starts an active communication with the network infrastructure, sending a so-called measurement report to the base station. This report consists of the received signal strength of up to six of the strongest neighboring cells and the received signal strength of the serving base station. The range between the periodic location updates may vary among the infrastructure providers.

3.1 Locating Mobile Phones

Due to regulatory requirements[5] but also driven by commercial opportunities locating mobile phones gained the attention of research and industry. There is a variety of possibilities for determining a mobile station's location from the view point of the infrastructure, e.g., by Cell Origin with timing advance (TA) and Uplink Time Difference of Arrival (U-TDOA) for GSM [21][6]. While the latter method requires sophisticated network infrastructure, Cell Origin and TA are available in any network setup. However, all these methods work without special requirements for the mobile station and achieve a positioning accuracy of up to 50 m in urban areas (TDOA) [23].

Another (nonstandard) method to determine a MS's location makes use of measurement results. Usually based on databases built from signal propagation models used during the planning phase of the infrastructure, this data can be used to create a look-up table for signal measurements to determine the MS's location. Based on the cell, TA and received signal strength of the serving cell as well as the six neighboring cells, Zimmermann et al. achieved positioning accuracy of below 80 m in 67% and 200 m in 95% in an urban scenario [24]. With a similar method but more generic setup, Peschke et al. report a positioning accuracy of 124 m in 67% [25]. While the mobile phone is in idle mode, network-assisted positioning is not possible. The network either has to wait for the next active period of the MS (e.g. phone call, location update) or has to initiate MS activity. This can be done by transmitting a *silent text message* to the MS in order to force an active communication, but without raising the user's awareness. The procedure is used for instance by law enforcement authorities or by location-based services based on GSM positioning.

3.2 Privacy Threats

Providing a proper (especially technical) definition of location privacy has proven to be a difficult task. Many different definitions were published, all covering

[5] e.g. FCC Enhanced 911 Wireless Service, `http://www.fcc.gov/pshs/services/911-services/enhanced911` [5/15/2011]

[6] Location determination options for UTRAN [22]

specific aspects. One abstract definition, first formulated by Westin [26] and modified by Duckham & Kulik [27], describes location privacy as:

> "[...] a special type of information privacy which concerns the claim of individuals to determine for themselves *when*, *how*, and to *what extent* location information about them is communicated to others."

According to this definition the user should be in control of the dissemination of his location information. Thus, the user's privacy is threatened (according to the aforementioned definition) because of technical necessities frequent location information is generated and possibly stored for different reasons (e.g., for technical network monitoring and improvement, regulation and law enforcement requirements). The user's mobile station collects and transmits location data without notification or consent. Furthermore, besides using the mobile device for active communication the user is unaware when location data is generated and transmitted (e.g. location updates, silent text messages, etc.). Furthermore, the user's privacy is threatened because of monetization of available location information. Even though this data is usually aggregated or anonymized, users are not aware of the technique used and thus bear a risk of re-identification (which they can't assess) (cf. [4]). Neither the final data consumer nor the intended use of the location data is known. Finally, users are not aware of the extent of the information they share. Usually one can assume that a single location datum does not reveal much information to a ubiquitous observer. In contrast to trusted communication peers such generic observers do not have appropriate background knowledge on a single individual and thus have difficulties deriving the user's real life context or current activity, especially for service providers with a subscriber base. However, by the accumulation of location observations knowledge about a user can be easily created. In order to improve the privacy situation in mobile communication networks, any location disclosure has to be made transparent and a privacy metric is required in order to evaluate the extent of location disclosures.

3.3 Adversary Model

In order to measure the extent of location disclosure in mobile communication networks and to assess the effects on the user's privacy the adversary has to be modeled. From a user's perspective, there is no assured knowledge on the capabilities of the observing / listening adversary, especially how disclosed or observed location data is used and what kind of conclusion the adversary is able to make based on the information gained. In general, the user's knowledge is limited to common knowledge about the technical and architectural characteristics of the mobile communication technology he or she uses (e.g. communication infrastructure service, etc.) as well as to a general estimation of the location determination abilities, limited by technical or physical factors. Furthermore, the user is able to monitor her own usage patterns by logging her exposure to the network, and has knowledge about the surrounding landscape, i.e., map knowledge.

Hence, the adversary model is limited to information an adversary may have gained during a defined observation period. We assume that an adversary A has a memory $O = \{o_1, \ldots, o_m\}$ of observations on the user's movement history based on time-stamped location observations $o_t = (c, \varepsilon)_t \in \mathbb{O}$, which are tuples of a geographic coordinate $c \in \mathbb{C}$ and a spatiotemporal error estimate $\varepsilon \in \mathbb{E}$ of this coordinate. The index t is a timestamp describing when the location observation was made, with o_m being the latest observation. The function $loc : \mathbb{O} \rightarrow \mathbb{C}$ extracts the location information from the tuple and $err : \mathbb{O} \rightarrow \mathbb{E}$ returns the error estimate.

In our scenario we assume that the adversary's utility (denoted as U_A) is negatively correlated with the user's privacy level in a communication relation with adversary A denoted as $P^A \in [0, -\infty)$, with $P^A = 0$ as the maximal achievable privacy level:

$$U_A(O) \simeq -P^A(O). \tag{1}$$

For instance if the user does not disclose any location information, her privacy is maximal but also the adversary's utility is zero. Henceforward, there is a utility gain if the adversary extends his knowledge either on the user's preferences or on his (periodic) behavior. This utility gain might be due to technical reasons (e.g. efficient network planning) or to the reuse of the data for commercial purposes. We can also assume that the adversary's utility is not decreased through any location data as long as the data is accurate, i.e., the user is not lying. Accordingly, $U_A(O') \geq U_A(O)$, with $O' := O \cup o'$, iff. o' reveals previously unknown information to the adversary A. Hence the user's privacy w.r.t. adversary A can only decrease by disclosing additional information: $P^A(O') \leq P^A(O)$.

Furthermore, the adversary's utility as well as the user's privacy depends on the nature and magnitude of the error estimate ε. First, with more accurate information possibly more information might be disclosed and thus, $err(o') < err(o) \Rightarrow U_A(o') \geq U_A(o)$ whereas the actual information gain is dependent, e.g., on landscape characteristics or additional knowledge on the user's context. Second, depending on the adversary and the kind of observation, the error value ε for a given location sample is evaluated differently. If the adversary determines the location by direct observation (o^{adv}), e.g., through WiFi/GSM/3G infrastructure, the adversary knows the size and distribution of the expected spatial error for the observed location sample. Furthermore, the temporal error component can be ignored. In contrast, if location information is given by the user (o^{usr}), the adversary has no information about the quality and thus the magnitude of the error ε of the observed sample. The user might have altered the spatial and/or temporal accuracy of the location information before submission. In general we can assume that $err(o^{adv}) \leq err(o^{usr})$ and therefore $U_A(o^{adv}) \geq U_A(o^{usr})$, since a robust error estimation reduces the adversary's uncertainty and thus increases the potential information gain. But more importantly the adversary chooses time and frequency of location observations.

3.4 Determining an Adversary's Knowledge Level

In order to reflect the duration, density, and quality of observation, a model of all past disclosures (i.e. history or knowledge (K)) w.r.t. a given adversary is required. The user's privacy is threatened by the discovery of his regular behavior and preferences (i.e. movement pattern). Since a user cannot change the knowledge an adversary already has, the user may evaluate the level of completeness of an adversary's information and the information gain or privacy loss involved in disclosing a further location sample.

Based on the adversary's utility function, we require that the knowledge gain $\Delta K_A(O, o') = K_A(O, o') - K_A(O \setminus \{o_m\}, o_m) \geq 0$ for any o'. If no new information is released, $\Delta K_A = 0$, and thus no privacy loss is experienced by the user. For a user it is important to know what extra information the disclosure of a single location sample o' gives to each listening or observing adversary A w.r.t. his history.

In a study on movement patterns of mobile phone users, Gonzalez et al. found a characteristic strong tendency of humans to return to places they visited before. Furthermore, the probability of returning to a location depends on the number of location samples for that location. A rough estimation can be denoted as $Pr(l_k) \sim k^{-1}$ where k is the rank of the location l based on the number of observations [1]. In a similar study it was shown that the range in the number of significant places is limited ($\approx$ 8-15). At these places a user spends about 85% of the time. However, there is a long tail area with several hundred places which were visited less than 1% but covered about 15% of the user's total observation time [7]. For the proposed privacy model we concentrate on the top L popular places (with L being in the range of about 8-15), as these places are likely to be revisited and therefore are considered as significant places in a user's routine.

If we assume that the attacker's a-priori knowledge on the observed location sample o' is limited to the generic probability distribution describing human mobility patterns and the accumulated knowledge so far, then we can model the adversary's knowledge as the uncertainty assigning the observed location information to a top L place. Entropy can be used to express the uncertainty of the adversary and therefore the user's privacy. Using entropy to quantify privacy was already used in different settings (e.g [28]). In the following we consider a location $l \in \mathbb{C}^*$ to be an arbitrarily shaped area in $\mathbb{C}$ and denote the spatial inclusion of a precise coordinate $c \in \mathbb{C}$ in the area l by writing $c \cong l$. To comply with the characteristics of human mobility patterns as described above, we define the probability of an observed location sample o' belonging to one of the top L locations ($l_i,\ i \in \{1, \ldots, L\}$) as $p_{l_i} := Pr(loc(o') \cong l_i) = \frac{\tau}{i}$ where $\tau \in (0, 1]$ is chosen in a way such that $\left(\sum_{i=1}^{L} p_{l_i}\right) + \gamma = 1$ with $\gamma \in [0, 1)$ representing the summed probability of o' belonging to one of the many seldom visited places in the long tail distribution observed by Bayir et al. [7]. Assuming that the adversary A has already discovered the top k locations of the user (by making use of the previously observed user locations in O), we make a distinction between two cases: (A) o' belongs to a frequently visited location already known to the adversary ($\exists i \in \{1, \ldots, k\} : loc(o') \cong l_i$), or (B) the adversary is not able

to unambiguously connect the location observation to an already detected top L location.

In case (A) no information about new frequently visited places is revealed (which we denote by $K_A^{L(A)}(O, o') = 0$). For case (B) we measure privacy as the uncertainty (i.e. entropy) of assigning o' to one of the remaining unknown top L locations. We denote with $p_{sk} := \sum_{i=1}^{k} p_{l_i}$ the summed probability for the k top locations *known* to the adversary and accordingly $p_{su} := \sum_{i=k+1}^{L} p_{l_i}$ the summed probability for the *unknown* top locations. Given that o' does not belong to one of the k known places, the probability for the remaining places $l_{k+1} \dots l_L$ changes to $p_{l_i}^k = p_{l_i} \cdot (1 + \frac{p_{sk}}{p_{su}})$, which yields the following entropy calculation:

$$K_A^{L(B)}(O, o') = -(\sum_{i=k+1}^{L} p_{l_i}^k \log p_{l_i}^k) - \gamma \log \gamma, \tag{2}$$

where γ denotes the summed probability of location samples which do not belong to the top L locations. The overall uncertainty level of the adversary is the weighted sum of the two cases (A) and (B) described above:

$$K_A^L(O, o') = p_{(A)} \cdot K_A^{L(A)}(O, o') + p_{(B)} \cdot K_A^{L(B)}(O, o'), \tag{3}$$

where $p_{(A)} = p_{sk}$ is the probability of case (A) and $p_{(B)} = 1 - p_{(A)}$ the probability of case (B). By merging the equations of the two cases, the overall uncertainty of an adversary in assigning o' to a yet unknown top location can be expressed as:

$$K_A^L(O, o') = (1 - p_{sk}) \cdot \left(-(\sum_{i=k+1}^{L} p_{l_i}^k \log p_{l_i}^k) - \gamma \log \gamma \right). \tag{4}$$

Until now, we assumed a simple binary decision as to whether a location sample belongs to a regular visited place (i.e. cluster) or not, hence $\varepsilon \simeq 0$ and a function $C_O(l) = |\{o \in O \mid loc(o) \cong l\}|$ counting the number of times a user was observed at a given location $l \in \mathbb{C}^*$ (cf. section 3.4 above), making it possible to rank the places by their popularity ($l_1, l_2, \dots$ where $C_O(l_i) \geq C_O(l_{i+1})$ – which means that l_1 is the most frequently visited location). Location information observed by mobile communication infrastructure is error prone. Depending on the communication infrastructure used, users can make assumptions on the physical limitations of the involved technology and thus can estimate a best case value for ε. In order to model the adversary's uncertainty we introduce p_c as the probability of function C^E assigning o' correctly to a location $l \in \mathbb{C}^*$, taking $\varepsilon = err(o')$ into account (and $\overline{p_c} := 1 - p_c$). As the precise definition of p_c depends on the implementation of C^E we only assume a correlation between the error and this probability: $p_c \sim \varepsilon^{-1}$.

However, modeling the adversary's uncertainty based on ε is in practice both difficult and possibly harmful to the user, since the adversary's capabilities might be underestimated, which may result in a higher and misleading privacy level. Hence, in order to get a robust reflection on a user's frequently visited places, using a clustering approach leads to an efficient but also abstract representation

of the user's regular behavior. Several studies (e.g. [15,29]) demonstrate that clustering is an effective tool for identification of a user's significant places.

Thus, instead of modeling the uncertainty of an adversary in assigning an observation to a certain location l_i, the spatial size of a possible location cluster is increased by the estimated spatial error. Thus, the adversary's uncertainty can be translated into the problem of choosing a single location out of all possible (and plausible) locations within the clustered spatial area. This uncertainty can be calculated using map data. Fig. 2 shows the resulting clusters for GPS data (left) and GSM data (right) from a 17-day trace with hourly location observations (GSM) and an estimated error of 250 m (GSM).

(a) (b)

Fig. 2. Clusters generated by a 17-day GPS trace (a) and (b) 17 days of hourly location updates (GSM) with an estimated spatial error of 250 m. The radius of each cluster denotes its significance for the user.

3.5 Determining an Adversary's Knowledge Gain

With the uncertainty value before and after disclosure of o', an adversary only gains new information if a new frequently visited location is uncovered and can be calculated as $\Delta K_A^L(O, o') = K_A^L(O, o') - K_A^L(O \setminus \{o_m\}, o_m)$ where o_m is the latest location observation in O (and therefore the direct predecessor of o').

If o' can be assigned to a known location $l_i \in L$, then $\Delta K_A^L = 0$, as by definition no information about new frequently visited places is revealed. However, the weight of already determined frequently visited places may change due to such an observation. Furthermore, people's preferences are not static and hence neither are their preferences as to frequently visited places. For instance, people

change employer and/or move from time to time. Such changes in regular behavior disclose private information and thus compromise the user's privacy. To model these changes, the observation horizon can be limited and any information older than a certain amount of time could be discarded.

To model changes in the frequency of the user's top locations and a user's regular behavior, we measure the change in the distribution made by a new observation. The adversary's a-priori knowledge is the distribution of the time spent in all known locations and hence their relative importance to the user. Thus, an adversary gains extra knowledge if the distribution of time spent changes, i.e., the user's preferences change. For every detected location we assume that the true probability $q(O, o', l_i) := Pr_O(loc(o') \cong l_i)$ is the relative observed importance of location l_i derived from the previous observations in O (e.g. $Pr_O(loc(o') \cong l_i) \sim C_O(l)$). We define the information gain as the difference between the observed distribution before and after a disclosure of additional data. One simple method for measuring the information gain is the relative entropy using KL-divergence [30]

$$K_A^C(O, o') = -\sum_{i=1}^{k} q(O, o', l_i) \log \frac{q(O, o', l_i)}{q((O \cup o'), o', l_i)} \quad , \tag{5}$$

where $q(O, o', l_i)$ denotes the probability of returning to l_i before and $q((O \cup o'), o', l_i)$ the new probability after the new observation o'.

Finally, we express the privacy loss as

$$\Delta K_A(O, o') = \Delta K_A^L(O, o') + K_A^C(O, o'). \tag{6}$$

4 Case Study GSM Network

In contrast to previous work with focus on analyzing call data records, our focus was on uncovering the side effects of using a mobile handset, since these are currently the most personal devices we know. The mobile handset is a highly sensitive device not only due to the huge subscriber count, but especially because people hardly do any activity without keeping their mobile phones nearby. The aforementioned studies showed the expressiveness of call data records. However, we are focusing especially on location updates, since these are scheduled periodically and configuration among network providers differs significantly. While one mobile telephony provider requires a client to initiate a LU every 60 minutes, another one only requires updates every 12 hours. The remaining two telephony providers configured their networks requiring four and six hour intervals respectively. From a user's privacy perspective, location updates are especially threatening because of their regularity but also because these events happen without the user noticing.

The information on the network infrastructure and its configuration a user gets through the handset's UI is usually limited to the mobile operator name, signal strength of the serving cell and type of network connection (e.g. GSM

or 3G). In order to analyze the user's exposure, we developed a logger device to record any communication between the GSM infrastructure and a mobile phone. This device was carried by test persons; however, the phone was kept in a passive / idle mode, i.e., no phone calls were made or received. The logger device is based on a Nokia 3310 phone. These phones are able to provide raw network data through a specific debug interface[7]. This data can be recorded, decoded and analyzed in a second step. To make this setup mobile, a micro controller writing the data to an SD-Card was attached. Furthermore, a GPS device was added to tag the network data with a time stamp and to record the user's movements. With this method we could not directly determine the knowledge the network infrastructure has. However, we could record each time the user was exposed to the infrastructure, and a network-based determination of his location was possible or location data was generated by the network infrastructure (e.g. measurement of the timing advance). In order to simulate the larger error of GSM positioning, a random radial error of ε was added to the GPS position. For our experiments we implemented a cluster algorithm based on a radius filter. For periodic and gap-based location data (e.g. GSM) such a filter simply reflects the frequency a user was observed at a specific place. Additionally, for GPS data a gap filter was used. With this method we were able to detect the places a user revisited frequently. Throughout the experiments a value of $\tau = 0.3$ was used, which roughly represents the results from the aforementioned studies on human mobility patterns. Furthermore, 12 clusters were expected. We assumed $\varepsilon = 250\,\mathrm{m}$ as the average positioning accuracy of the GSM network.

4.1 Data Analysis

The first analyzed data set was created by a test person carrying the logger device for about 17 days equipped with a SIM-card from a German provider which requires location updates hourly. 17744 GPS points and 312 location updates were recorded. The reason that the number of location updates is lower than anticipated is twofold: the first and the last day were not complete, but there were also signal losses and user operation errors like empty batteries. However, these results should correspond with real life mobile phone usage. During that time a total of 10 clusters could be identified through the GPS data, 8 based on GSM data. The remaining clusters found by the GPS method were not detected. This is due not only to the short evaluation period and lower spatial resolution of the GSM positioning but also to the short length of stay at the remaining two places (i.e. less than one hour). In a test trial with 6 hourly location updates and an equal test period, only three clusters could be detected. Fig. 3a shows the temporal development of the discovery of frequently visited places using different methods and location sample frequency. Fig. 3b shows the adversary's knowledge on the remaining, yet uncovered, places. In a further trial with 12-hour location update intervals, within 8 days only a single cluster could

[7] GSM decoding with a Nokia 3310 phone, `https://svn.berlin.ccc.de/projects/airprobe/wiki/tracelog`, [5/15/2011]

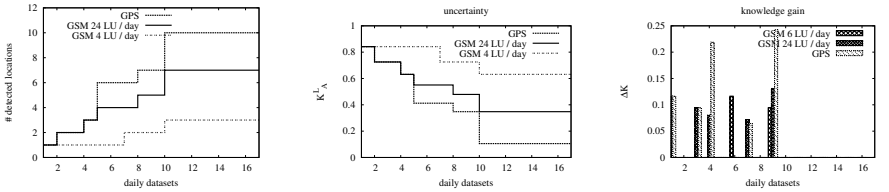

Fig. 3. Modeling an adversary's knowledge gain using GSM and GPS data sources. (a) shows the temporal development of detecting frequently visited places, (b) shows the uncertainty assigning a new observation to a yet unknown frequently visited place and (c) shows the adversary's knowledge gain (ΔK) by disclosing a daily data set.

be determined. One reason was the disadvantageous time points at 7:30 AM and 7:30 PM. However, the time points were chosen by the network. For this configuration a long term trial is pending. Due to long distance traveling, offline phases, and time periods without reception, we expect random shifts for the time point of location updates. Therefore, for a long term observation it seems likely that some (2-3) additional clusters should be detected. The probability of detecting a cluster where a user spends large amounts of time is more likely and thus is likely to be uncovered first. The privacy measurement also implicitly captures the distribution of the observed location samples. If the distribution of location samples is concentrated within certain time spans, fewer clusters will be discovered. The same applies for evenly distributed but sparse samples (e.g. every 12 hours).

4.2 Privacy Improvements

Based on the aforementioned analysis, several enhancements could improve the user's privacy in mobile communication networks. First, one can observe that a simple quantitative privacy policy as offered by network providers, stating only the length of possible data storage is neither meaningful nor helpful for a subscriber w.r.t. location privacy. Especially the density of periodic location samples makes a significant difference as to the provider's possible knowledge base and thus the user's present and future privacy risks. Therefore, subscribers also need to know when, how and to what extent location information is generated. With such knowledge the user's awareness as to his privacy loss is raised. In a second step the user should be able to control location dissemination by making informed decisions.

A *privacy aware* mobile phone requires software interfaces to the mobile phone stack controlling and exposing signaling attempts (e.g. detecting silent text messages), measurement reports and the occurrence of location updates. With the help of *osmoconbb* GSM baseband implementations, first steps toward a privacy aware phone were made. The mobile station is able to log location data and expose it to the user, which is sent to the service provider The measurement

results sent by the MS during location updates include signal strength measurements from surrounding BTS. The measurement information is used for the handover decision during the connection. Since a LU requires only a very brief communication with the network, a handover between different cells is very unlikely or even impossible. Thus, sending measurements of neighboring stations is technically not always required. If the number of transmitted measurements is reduced or completely omitted, the accuracy of the network's position estimation is significantly decreased. In the best case (no measurements transmitted), the accuracy is decreased to cell origin with timing advance. A further step to decrease the accuracy of the position determination is transmitting modified or false measurement information. To decrease the accuracy of the position estimation further, a MS could send with a slight timing offset. Such offsets have direct impact on the timing advance calculation of the BTS. Consequently, this leads to an incorrect distance estimation between MS and BTS. The combination of manipulating measurement results and timing advance gives the possibility to conceal the actual position of the MS. The rough position of the MS is still given by the BTS used and its covered area.

5 Conclusion

We fully acknowledge that the evaluation is based on too little data to be statistically significant. However, the data clearly indicates that network configuration has an impact on the user's location privacy. Furthermore, the proposed metric gives the user a tool to understand the impact of mobile communication on his privacy without knowing the adversary's capabilities or behavior.

In contrast to other personal digital devices, mobile phones are hardly ever switched off, thus offering unique options for (unobserved) user tracking. The disclosure or detection of any significant place decreases the user's privacy by roughly the same level, independent of the relative importance or rank of the place. The proposed user model and accordingly the privacy metric showed that the user's privacy loss is roughly the same for all detected clusters. Especially for a setting with semi-trusted adversaries, this result reflects the (commercial) importance of lower ranked clusters w.r.t. the completeness of a user's profile. Since lower ranked clusters are harder to detect, the ability to uncover such a place reflects the density and/or the length of observation by an adversary and thus on the user's exposure. Therefore, network configuration is crucial for the individual's privacy. In our tests we saw time ranges for location updates range from 60 minutes to 12 hours with different results for detecting frequently visited places. These different network configurations have a significant impact on the user's location privacy especially if privacy policies only specify the duration of data storage. Finally, countermeasures to improve the user's privacy were proposed. The evaluation of the effectiveness of the proposed actions remains for future work.

References

1. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature 453, 779–782 (2008)
2. Krumm, J.: Ubiquitous advertising: The killer application for the 21st century. IEEE Pervasive Computing 99 (2010)
3. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: MobiSys 2003: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, pp. 31–42. ACM, New York (2003)
4. Ma, C.Y., Yau, D.K., Yip, N.K., Rao, N.S.: Privacy vulnerability of published anonymous mobility traces. In: Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking, MobiCom 2010, pp. 185–196. ACM, New York (2010)
5. Bettini, C., Wang, X.S., Jajodia, S.: Protecting privacy against location-based personal identification. In: Jonker, W., Petković, M. (eds.) SDM 2005. LNCS, vol. 3674, pp. 185–199. Springer, Heidelberg (2005)
6. Golle, P., Partridge, K.: On the anonymity of home/Work location pairs. In: Tokuda, H., Beigl, M., Friday, A., Brush, A., Tobe, Y. (eds.) Pervasive 2009. LNCS, vol. 5538, pp. 390–397. Springer, Heidelberg (2009)
7. Bayir, M., Demirbas, M., Eagle, N.: Discovering spatiotemporal mobility profiles of cellphone users. In: IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops, WoWMoM 2004, pp. 1–9 (2009)
8. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Rowland, J., Varshavsky, A.: A tale of two cities. In: HotMobile 2010: Proceedings of the Eleventh Workshop on Mobile Computing Systems; Applications, pp. 19–24. ACM, New York (2010)
9. Girardin, F., Calabrese, F., Dal Fiorre, F., Biderman, A., Ratti, C., Blat, J.: Uncovering the presence and movements of tourists from user-generated content. In: Proceedings of International Forum on Tourism Statistics (2008)
10. Girardin, F., Vaccari, A., Gerber, A., Biderman, A., Ratti, C.: Towards estimating the presence of visitors from the aggragate mobile phone network activity they generate. In: Proceedings of International Conference on Computers in Urban Planning and Urban Management (2009)
11. Sohn, T., Varshavsky, A., LaMarca, A., Chen, M., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W., de Lara, E.: Mobility detection using everyday GSM traces. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 212–224. Springer, Heidelberg (2006)
12. De Mulder, Y., Danezis, G., Batina, L., Preneel, B.: Identification via location-profiling in gsm networks. In: Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society, WPES 2008, pp. 23–32. ACM, New York (2008)
13. Lee, C.H., Hwang, M.S., Yang, W.P.: Enhanced privacy and authentication for the global system for mobile communications. Wirel. Netw. 5, 231–243 (1999)
14. Ardagna, C., Jajodia, S., Samarati, P., Stavrou, A.: Privacy preservation over untrusted mobile networks. In: Bettini, C., Jajodia, S., Samarati, P., Wang, X. (eds.) Privacy in Location-Based Applications. LNCS, vol. 5599, pp. 84–105. Springer, Heidelberg (2009)
15. Hoh, B., Gruteser, M., Xiong, H., Alrabady, A.: Achieving guaranteed anonymity in gps traces via uncertainty-aware path cloaking. IEEE Transactions on Mobile Computing 9, 1089–1107 (2010)

16. Shokri, R., Freudiger, J., Jadliwala, M., Hubaux, J.P.: A distortion-based metric for location privacy. In: WPES 2009: Proceedings of the 8th ACM Workshop on Privacy in the Electronic Society, pp. 21–30. ACM, New York (2009)
17. Beresford, A., Stajano, F.: Location privacy in pervasive computing. IEEE Pervasive Computing 2, 46–55 (2003)
18. Duckham, M., Kulik, L.: A formal model of obfuscation and negotiation for location privacy. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 152–170. Springer, Heidelberg (2005)
19. 3rd Generation Partnership Project (3GPP): TS 24.008 Technical Specification Group Core Network and Terminals; Mobile radio interface Layer 3 specification; Core network protocols; Stage 3 (Release 10) (2010)
20. 3rd Generation Partnership Project (3GPP): TS 45.008 Technical Specification Group GSM/EDGE Radio Access Network; Radio subsystem link control (Release 9) (2010)
21. 3rd Generation Partnership Project (3GPP): TS 43.059 Technical Specification Group GSM/EDGE Radio Access Network; Functional stage 2 description of Location Services (LCS) in GERAN (Release 9) (2009)
22. 3rd Generation Partnership Project (3GPP): TS 25.305 Technical Specification Group Radio Access Network; Stage 2 functional specification of User Equipment (UE) positioning in UTRAN (Release 10) (2010)
23. Sun, G., Chen, J., Guo, W., Liu, K.: Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs. IEEE Signal Processing Magazine 22, 12–23 (2005)
24. Zimmermann, D., Baumann, J., Layh, A., Landstorfer, F., Hoppe, R., Wolfle, G.: Database correlation for positioning of mobile terminals in cellular networks using wave propagation models. In: IEEE 60th Vehicular Technology Conference, VTC 2004-Fall, vol. 7, pp. 4682–4686 (2004)
25. Haeb-Umbach, R., Peschke, S.: A novel similarity measure for positioning cellular phones by a comparison with a database of signal power levels, vol. 56, pp. 368–372 (2007)
26. Westin, A.F.: Privacy and Freedom, 1st edn. Atheneum, New York (1967)
27. Duckham, M., Kulik, L.: Location privacy and location-aware computing, pp. 35–51. CRC Press, Boca Raton (2006)
28. Díaz, C., Seys, S., Claessens, J., Preneel, B.: Towards measuring anonymity. In: Dingledine, R., Syverson, P.F. (eds.) PET 2002. LNCS, vol. 2482, pp. 54–68. Springer, Heidelberg (2003)
29. Ashbrook, D., Starner, T.: Using gps to learn significant locations and predict movement across multiple users. Personal Ubiquitous Comput. 7, 275–286 (2003)
30. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22, 79–86 (1951)