Location Privacy in Relation to Trusted Peers

Klaus Rechert[1] and Benjamin Greschbach[2]

[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. One common assumption when defining location privacy metrics is that one is dealing with attackers who have the objective of re-identifying an individual out of an anonymized data set. However, in today's communication scenarios, user communication and information exchange with (partially) trusted peers is very common, e.g., in communication via social applications. When disclosing voluntarily a single observation to a (partially) trusted communication peer, the user's privacy seems to be unharmed. However, location data is able to transport much more information than the simple fact of a user being at a specific location. Hence, a user-centric privacy metric is required in order to measure the extent of exposure by releasing (a set of) location observations. The goal of such a metric is to enable individuals to estimate the privacy loss caused by disclosing further location information in a specific communication scenario and thus enabling the user to make informed choices, e.g., choose the right protection mechanism.

1 Introduction

Location information has recently become a popular but also valuable communication item. Ubiquitous and affordable mobile communication paired with a new generation of so-called Smartphones has given rise to a large variety of location based applications. However, exploitation of mobile location information also brings new challenges to the users' privacy.

Providing a proper definition of location privacy has proven to be a difficult task. Many different definitions were published, all covering specific aspects. One abstract definition, first defined by Westin [1] and modified by Duckham & Kulik [2], 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. Location sharing usually involves location data as coordinates related to a sphere or map. Depending on the source, this information might be error prone. For instance, the accuracy of GPS location determination using a consumer device (Smartphone) might range from 1 to 50 meters; location

C. Meadows and C. Fernández-Gago (Eds.): STM 2011, LNCS 7170, pp. 106–121, 2012.

determination utilizing a GSM/3G infrastructure might have an error range of 50-250 meters.

In today's communication scenarios user communication and information exchange with (partially) trusted peers is very common, e.g., in communication via social applications. In disclosing voluntarily a single observation to a (partially) trusted communication peer, the user's privacy seems to be unharmed (using the aforementioned definition). However, location data is able to transport much more information than the simple fact of a user being at a specific location. In the long run, location data is able to describe what a user has done and what he is currently doing.

For instance, a single location observation might have a different impact on the user's privacy depending on time and place but also on the *observer*. The observer might be able to make exact conclusions about the user's state and intention, if the observer has good background knowledge about the user (e.g., wife, friends). Even observers with little or no background knowledge are able to gain knowledge about the user. For example, by observing a user's frequently visited places, one can make conclusions about the user's workplace or other preferences. Using Westin's definition, it is difficult for a user to measure the *extent* of his location disclosure, especially with trusted communication peers where an anonymity approach is unsuitable. Hence, a user-centric privacy metric is required to measure the extent of exposure caused by releasing (a set of) location observations. From a user's perspective, with the goal of minimizing exposure, only as little information as possible should be disclosed.

2 Related Work

Privacy metrics is an important field in research on mobile communication and location based services, since they provide the fundamental model to evaluate a privacy protecting scheme. One way of characterizing a (location) privacy metric is the underlying adversary model: the metric describes how successfully one's privacy is protected against the defined adversary.

A popular model is an adversary that observes in some way generalized location data and tries to reconstruct this data based on 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. [3]). 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 [4]. A popular privacy metric is *k-anonymity*, developed in [5] and further extended for a location context (e.g. [6,7]). A single variable is able to determine a user's privacy level, i.e., being indistinguishable from $k-1$ other agents. However, this metric may be misleading if all k users are within a region with only a few plausible positions. l-diversity and road segment s-diversity avoid this issue by only taking plausible positions into account [8]. Furthermore, k-anonymity and similar methods imply that a suitable number of cooperative agents are available for a specific service or listening group and global knowledge about the state of other agents is required. Thus, a user cannot determine or preserve a desired privacy level in an autonomous manner. Neither does this metric cover the sensitivity of a location at a given time [4], nor is it able to fully protect specific movement patterns [9].

A different method for measuring location privacy is to make use of the uncertainty of an adversary 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 [10]. 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 [3].

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, and they usually are based on the assumption that the user requires full anonymity. Hence, such measures are not suitable for communication with (semi-)trusted peers (e.g., social contacts) or in ubiquitous communication networks which require a confirmed user identity. Furthermore, such models assume that for every available service there is a sufficient number of cooperative agents nearby and such an approach is usually applicable for a subset of location based service.

Cranshaw et al. developed an entropy-based approach for analyzing the social context of a geographic region. The proposed model assigns a high entropy to a place if a large variety of users was observed at that location, a low entropy value if the place was visited by only a few users [11]. Based on the location diversity measurement above, a user-study was conducted on presence-sharing preferences. Toch et al. found that people are more comfortable sharing their location in places which are visited by a large and diverse group of people in contrast to places which are highly frequented but by a homogeneous group [12]. Diaz et al. introduced a measure of entropy to quantify the degree of anonymity a mix-network provides [13]. Kamiyama et al. extended the entropy measure to quantify information disclosure through various media [14]. The described measurement quantifies the privacy loss caused by the disclosure of several (sensitive) attributes.

3 User-Centric Location Privacy Metric

For a user-centric location privacy model, location privacy has to be seen from a different angle. As the user is not always able to hide or remain anonymous, she could still achieve insight on the possible knowledge base of the communication peers involved and thus could achieve or increase location privacy (w.r.t. the aforementioned definition) through informed decisions on *when*, *how* and to *what extent* she discloses her location information. Hence, an evaluation of the user's location in the context of each listener group is necessary.

3.1 Adversary Model

In terms of (location) privacy, all communication peers are considered only as partially trusted, because once data is exchanged, this information usually cannot be recalled by the user. Even when considering explicit (legal contract) or implicit (social contract) based privacy policies, the control problem remains. Hence, all partially trusted peers are also considered as adversaries. Furthermore, from a user's perspective, there is no certain 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 collected. Hence, the adversary model is limited to information an adversary may have collected 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 an 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. The choice of the geographic coordinate system ($\mathbb{C}$) and the concrete representation of the error ($\mathbb{E}$) is not important in the context of this paper.

In our scenario the user's utility is positive in a communication relation with communication peer (adversary) A. Otherwise a rational agent would not share information. We make a similar assumption for the adversary's utility ($U_A(o_t) \geq 0$). A separation of the user's utility disclosing information and the user's level of privacy is required, as the utility of location information naturally conflicts with the user's privacy level. In order to benefit from location-aware services, the user's location disclosure is required. Thus, for any location disclosure the user's privacy might decrease. Hence, the adversary's utility 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, the user's privacy is maximal but the adversary's utility is zero. Thereafter there is a utility gain if the adversary extends his knowledge either on the user's preferences or on his (periodic) behavior. 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: $P^A(O') \leq P^A(O)$.

An increase in the user's privacy level is only possible if the user is intentionally lying about his location, because providing false information may degrade the adversary's knowledge base or may lead to false conclusions. However, by providing false location information the user's utility decreases as well. For instance, in the case of location-based services a decrease in the user's utility might be caused by a decreased quality of service. In a communication scenario with social contacts, getting caught lying might lead to negative social consequences. For the rest of this paper we therefore assume that location observations reflect the true positions of the user.

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 more information might possibly 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 and application characteristics. Second, the error value ε for a given location sample is evaluated differently depending on the adversary and the kind of observation. If the adversary determines the location by direct observation (o^{adv}), e.g., through a WiFi/GSM/3G infrastructure, the adversary knows the size and distribution of the expected error for the observed location sample. 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 for the same given error ε. But more importantly the adversary chooses time and frequency of location observations.

3.2 Measuring Location Privacy

To measure the user's privacy or privacy loss, the objective measurable components defining P^A w.r.t. location observations have to be identified. Taking into account the aforementioned privacy definition and adversary model, the evaluation of observations regarding new information about the user is required. This information can be split into two parts: (1) gaining *knowledge* on the user's regular behavior and preferences (e.g., his neighborhood, occupation, leisure activities or social contacts) and (2) deriving *sensitive* private information on his current context (e.g., his activity or intention at an observed place).

We define the change of the user's privacy level due to a new location observation o' made by an adversary A who already has a location record O about the user straightforwardly by $\Delta P^A(O, o') := P^A(O \cup o') - P^A(O)$. According to the requirements from the adversary model with $U_A(O') \geq U_A(O)$ and $U_A \simeq -P^A$ it follows that $\Delta P^A(O, o') \leq 0$.

Knowledge. In order to reflect the duration, density and quality of an observation, a model of all past disclosures, i.e., history or knowledge K, to 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 as well as privacy loss for disclosing a further location sample.

Based on the adversary's utility function, we require that $\Delta K(O, o') = K(O \cup o') - K(O) \geq 0$. If no new information is released, $\Delta K = 0$ and thus no privacy loss is experienced by the user. Section 4 presents an example implementation of K.

Sensitivity. The second component threatening the user's privacy w.r.t. his current location is the sensitivity S of an observation o_t. Due to diverse preferences the individual subjective sensitivity of a certain location cannot be expressed in a generic way. However, an objective measure of location sensitivity is the level of the potential exposure caused by disclosing the user's location at a given time and date. The user is *exposing* himself by allowing or providing location observations. As in daily life, such behavior may provide new, possibly sensitive knowledge to any observer. However, in a crowded shopping or business district during business hours the user's exposure is limited. Even with knowledge of his current location, the user is hard to spot and therefore it is hard to observe his current activities or guess the user's intention, because the number and diversity of possible places where a user could be are rather high.

Similarly, S describes how difficult it is for an observer to observe or derive the user's real-life activity for a given (set of) location observation(s). Note that the observer may have good background knowledge about the user and therefore be able to derive the user's activity with little or rough and error prone location data. More formally $S(O, o_t) \in [0, 1]$ expresses the probability of an adversary being able to derive the current activity of the user, i.e., the reason for her visiting location $loc(o_t)$, taking previously visited locations O into account (especially the latest, $o_m \in O$). For $S(O, o_t) = 0$ the adversary does not learn anything about the user's activity or motivation for being located at o_t, while in the case of $S(O, o_t) = 1$ the adversary can derive this information from the location data without any doubts. Due to the spatiotemporal error $\varepsilon \in \mathbb{E}$, o_t describes only an area in $\mathbb{C}$ where the user might be located. Let c'_t be the actual precise location of the user at time t. If c^{ae}_t is the adversary's estimate of the location of the user at time t (making use of background knowledge of the user and external map knowledge), then $S(O, o_t) = Pr(c'_t = c^{ae}_t)$. Hence, with a growing spatiotemporal error and/or a dense and diverse landscape, the number of possible locations where a user could be increases and thus also the adversary's uncertainty regarding the user's action.

In section 5 an example implementation of location sensitivity is discussed.

Trust Relation. Third, the level of trust (denoted as θ) for a given adversary has to be modeled. In our communication scenario, the level of trust is defined as the estimated personal background knowledge a specific adversary already has about the user, based on the assumption that the user has trusted a peer to a certain extent, such that he has previously disclosed a certain amount of personal information, possibly through a different channel.

For instance, while communicating with social peers θ is more important, as with growing personal trust social contacts already have a good knowledge from other sources than mobile or social applications of the user's behavior in particular. Hence, the sensitivity of the current location might cause the individual to be more exposed, e.g., it might trigger uncomfortable questions, since these peers are able to infer subjectively sensitive places by using their background knowledge. In a communication relation with less trusted adversaries, e.g., location based services without (or with pseudonymous) registration, the protection of the user's daily routines is more important as there is usually little or no personal background knowledge. By disclosing regular patterns, the user might be identified (cf. [15,16]). By contrast, a single location sample without context on the observed individual has only little or no information value regarding the user's preferences or habits. The value of θ can be either predefined per classification of the listener class A or can be used as a user-parameter.

Definition of Privacy Loss. To formalize the discussion above, the privacy loss ΔP^A w.r.t. an adversary A, a set of m past location observations O of this adversary and a new location sample o' is

$$-\Delta P^A(O,o') = (1-\theta_A)\Delta K_A(O,o') + \theta_A S_A(O,o') \tag{2}$$

which is the weighted knowledge gain on the user's preferences ΔK_A and the location sensitivity S_A. This proposed location privacy metric captures the relative privacy loss, instead of measuring a privacy level. Especially in environments with (partially) trusted peers, the comparison of privacy levels is difficult because of the different relations and knowledge between users. By measuring only the relative privacy loss, different adversaries can be compared. Furthermore, the sensitivity measure is bound to a certain context. Thus, there is no absolute level of location sensitivity over time.

Comparison with Anonymity Metrics. In the case of a full anonymity scenario, we assume no trust relation at all to be existent between the user and the observer. Therefore we expect no background knowledge about the user on the observer's side and choose $\theta = 0$ accordingly. That implies that only the level of K matters for the privacy (or anonymity) level. By definition K describes the length, density and quality of the adversary's observations. In the case of an anonymity metric it describes the length of the observation of a single pseudonym and the level of knowledge gained about the user by observation. Thus, for any $\Delta K > 0$ the probability of being anonymous decreases. For instance, a simple user-centric estimation on the level of anonymity could be calculated based on the results by Golle and Partridge [15].

4 Example Implementation of K

In order to calculate the user's privacy level the adversary's knowledge (gain) has to be modeled. We assumed a knowledge gain / privacy loss only if the adversary learns some previously unknown information. For a user it is important to know what extra information the disclosure of a single location sample o' gives to an adversary A w.r.t. the adversary's observation 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 [17]. In a similar study it was shown that the number of significant places is limited ($\approx$ 8-15). A user spends about 85% of the time at these places. However, there is a long tail area with several hundred places which are visited less than 1% of the time but make up about 15% of the user's total observation time [18]. 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 about 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. 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 area l by writing $c \cong l$. To comply with the characteristics of human mobility patterns, 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 for o' belonging to one of the many seldom visited places in the long tail distribution observed by Bayir et al. [18]. 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. For case (B) we measure privacy as the uncertainty (i.e., entropy) on 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 $l_{k+1} \ldots l_L$ changes to $p_{l_i}^k = p_{l_i} \cdot (1 + \frac{p_{sk}}{p_{su}})$, which yields the following entropy calculation:

$$K_A^{L(B)}(O, o') = -(\sum_{i=k+1}^{L} p_{l_i}^k \log p_{l_i}^k) - \gamma \log \gamma \quad , \tag{3}$$

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:

$$K_A^L(O, o') = p_{(A)} \cdot K_A^{L(A)}(O, o') + p_{(B)} \cdot K_A^{L(B)}(O, o') \quad , \tag{4}$$

where $p_{(A)} = p_{sk}$ is the probability of case (A) and $p_{(B)} = 1 - p_{(A)}$ the probability of case (B). Integrating the two formulas of the two cases, the overall uncertainty of an adversary for connecting o' with a top location is:

$$K_A^L(O, o') = (1 - p_{sk}) \cdot \left(-(\sum_{i=k+1}^{L} p_{l_i}^k \log p_{l_i}^k) - \gamma \log \gamma \right) \quad . \tag{5}$$

4.1 Uncertainty of a Location Observation

In order to get a robust reflection of 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. [3,19]) have demonstrated that clustering is an effective tool for identification of a user's significant places.

However, the estimated or given horizontal positioning error has to be taken into account. Location information is usually expressed as inaccurate data, regardless of the error source, which is either data degraded on purpose or due to technical issues like an error prone positioning determination. Until now, we assumed a simple binary decision as to whether a location sample belongs to a regularly visited place (i.e., cluster) or not, hence $\varepsilon \cong 0$ 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 $l \in \mathbb{C}^*$ (see section 4 above), making it possible to rank the places by their popularity ($l_1, l_2, \ldots l_L$, with $C_O(l_i) \geq C_O(l_{i+1})$ – which means that l_1 is the most frequently visited location). In a more realistic setting location information is error prone. Depending on the nature of the observation, the effect of $\varepsilon > 0$ is different. If the user performs the location determination, the estimated error based on the technology used is known to the user but not to the adversary. Furthermore, users might deliberately increase ε to protect their privacy.

If the location is directly observed by the adversary, both user and adversary have knowledge on the possible error distribution depending on the technology used. Depending on the communication infrastructure used, users can make assumptions about the physical limitations of the technology involved 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: $p_c \sim \varepsilon^{-1}$.

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, resulting in a higher and misleading privacy level. Due to limited user knowledge, a default value of $\varepsilon = 0$ is used to simulate worst case knowledge and to avoid a possibly dangerous false sense of privacy. Still, ε remains an optional variable to the user.

4.2 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 can change due to such an observation. People's preferences are not static and hence neither are their preferences regarding frequently visited places. For instance, people change employer (or workplace) and/or move from time to time. Such changes in regular behavior cause private information to be disclosed and thus harm the user's privacy. To model these changes, the observation horizon can be limited and any information older than a certain amount of time is discarded.

To model changes in the frequency of the user's top locations and a user's regular behavior, we measure the change in distribution made by a new observation. The adversary's a-priori knowledge is the distribution of the time spent at all known locations and hence their relative importance to the user. Thus, an adversary gains extra knowledge if the distribution of time spent has changed, i.e., the user's preferences have changed. 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 the disclosure of additional data. One simple method to measure the information gain is the relative entropy using KL-divergence [20]

$$K_A^C(O,o') = -\sum_{i=1}^{k} q(O,o',l_i) \log \frac{q(O,o',l_i)}{q((O \cup o'),o',l_i)} \quad , \tag{6}$$

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

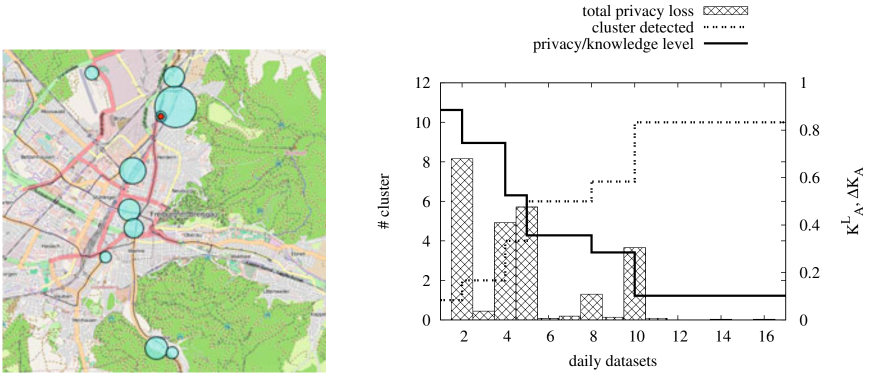
$$\Delta K_A(O,o') = \Delta K_A^L(O,o') + K_A^C(O,o'). \tag{7}$$

The privacy metric component ΔK_A, expressing knowledge about the user's preferred places, only measures the relative distribution of the times a user was observed at a specific place. Thus it is applicable for location based services without continuous observation or traces (e.g., location updates through an SNS are usually not continuous traces and appear infrequently).

4.3 Example

For our experiments we implemented a location cluster function based on a radius filter. For periodic and gap based location data (e.g., GSM) such a filter simply reflects how often a user was observed at a specific place. Additionally, for GPS data a gap filter was used to cover periods without GPS reception. 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. Figure 1(a) and 1(b) show 10 detected clusters from a 17-day GPS trace from a single user with a total of 17744 recorded GPS points. To measure the knowledge gain the data was segmented into daily data sets. After about 11 days the values of the final result of 10 clusters were discovered and remained constant afterwards.

The user's privacy level, based on the K_A^L, decreases almost linearly with the detection or disclosure of regularly visited places. The privacy loss caused by disclosing a low ranked place and thus with a low probability of being revisited is almost equal to that caused by disclosing a high ranked place. 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, uncovering such a place reflects the density and/or the length of observation of an adversary and thus the user's exposure.



(a)

(b)

Fig. 1. A user's privacy profile to a single adversary A based on a 17-day GPS trace. The data for K_A^L was normalized to 1.0. (a) shows the cluster result after 17 days; radius of each cluster denotes its relative importance; (b) shows detected clusters and calculated privacy gain/level.

5 Example Implementation of *S*

The last component of a user-centric privacy metric is the sensitivity of a given location and time. In contrast to the knowledge about the user's regular behavior which an adversary could extract from frequently visited places, users might evaluate the sensitivity of certain locations w.r.t. location privacy differently at different times, depending both on the type of place and the actual listener group.

However, the sensitivity of a location at a certain time can only be measured in objective terms. Personal preferences are too diverse and a generic formalization of possible subjective measures is difficult.

5.1 Static Location Sensitivity

Based on the ideas of location l-diversity [8] and related variants we define the sensitivity of a location $l \in \mathbb{C}^*$ as the user's plausible deniability of being at a (possible) subjectively sensible location $\hat{l}$, w.r.t. the knowledge of time t and the estimated location error ε. This definition can also be rewritten as the probability of an individual being at location l^* but observed at location l. Thereby l^* is an alternative plausible location in $\mathbb{C}^*$, which is not considered as subjectively sensitive.

However, taking only into account the number of plausible positions is not always sufficient. The number, distribution and especially the nature of the possible locations matter as well. For instance, if a person is in an area with a high density of landmarks (points of interest (POIs)), an adversary's uncertainty is high regarding the user's motivation in visiting the observed area. Furthermore, with a greater number of people nearby or visiting an area, a user's privacy increases (cf. [12]). Therefore, a discounting factor $\rho \in (0..1]$ is introduced, describing the nature of a given area, i.e., decreasing the "plausibility" depending on the listener and/or time of day. The static location sensitivity is defined as

$$S_A^S(o_t) = \frac{1}{\mathrm{numloc}(o_t)\rho(o_t)}. \tag{8}$$

The size of the area is defined by the maximum possible horizontal (deliberate or technical) location error ε and the maximum velocity at which a user can move. Function $numloc : \mathbb{O} \rightarrow \mathbb{N}$ returns the number of plausible positions for a given location sample, based on map-data. While *numloc* is a static measurement (i.e., the geographic features are considered static), ρ is time dependent, because the use cases for the landscape change depending, for example, on the time of the day, the day of the week, the season, etc..

5.2 Dynamic Location Sensitivity

A static measurement only captures an isolated observation. In most cases people move and submit their location continuously or frequently. Therefore, the sensitivity evaluation should also contain a dynamic, time-dependent component. For instance, the adversary only knows the published positions but not the exact route in between. The adversary may use a routing algorithm to determine a likely route a user could have taken. If there is only a single route, the adversary gains perfect knowledge. Consequently, the user gains privacy if the ambiguity of possible routes increases and thus the uncertainty of the adversary regarding the locations where a user could have been between two consecutive location disclosures.

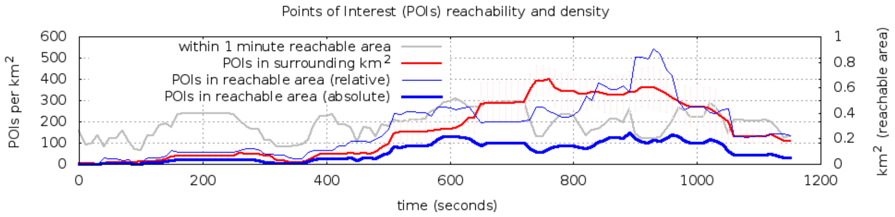


Fig. 2. A user trace of 1160 seconds, traveling 8.24 km showing the number of plausible locations and reachable POIs within one minute

We extend the static definition by including the time frame between two consecutive location disclosures as

$$S_A^D(O, o_t) = \frac{1}{\text{numreach}(o_m, o_t)\rho(o_m, o_t)}, \tag{9}$$

with a function *numreach* : $\mathbb{O} \times \mathbb{O} \to \mathbb{N}$ calculating the number of plausible locations in the reachable area between two consecutive location disclosures $o_m \in O$ (the latest in O) and o_t, and at a given velocity. Thus, the sensitivity component reflects the objectively measurable sensitivity of the user's current position by incorporating accuracy of the location determination, time, density of measurements and landscape.

5.3 Example

The sensitivity metric measures the information gain of an adversary knowing the user's current location. In contrast to the metric on cyclic behavior, we assume the adversary's information gain is derived from direct inference of the current location and the incorporation of external knowledge (e.g., map data). For our experiments we used data from OpenStreetmap (OSM). [1] The project provides accurate and deep map data for the evaluated region and allows the characterization of possibly sensible locations (e.g., public buildings, medical facilities, banks, etc. are marked through various attributes). Furthermore, the data can be downloaded and stored on a mobile device in order to make autonomous decisions without network access.

Two simple example components of calculating ρ by exploiting map features are the density of reachable landmarks or POIs and the expected population density for a given location and time. Figure 2 shows a sample trace of about 19 minutes traveling 8.24 km through the city. The trace was started in a business/industrial area, went through a residential area (around 150 - 250 and from 500) before entering the city center (around 750 - 1000). While the number of reachable plausible positions remains roughly at the same level, the number of reachable POIs increases in the city center significantly.

[1] The OpenStreetmap Project, `http://openstreetmap.org`, [1/15/2011].

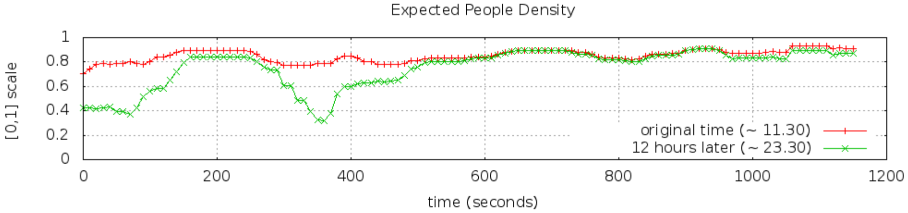


Fig. 3. Day-time dependent expected person density for a user trace of 1160 seconds, traveling 8.24 km. Calculation based on area classification based on OpenStreetmap data.

As a second example for the same trace, the expected person density was calculated. For each OSM *landuse*-tag[2] attribute a non-empirical estimation of expected person density was made. For instance, residential areas were assigned a high value for every time of day; for commercial and industrial areas a high value during business hours but otherwise a low value seems appropriate. For future work, empirical values need to be adopted. Figure 3 shows that during the day there is little variation, basically due to the fact that the trace never left city boundaries. However, at night there is a noticeable drop while crossing a business/industrial area (to 150 from 350).

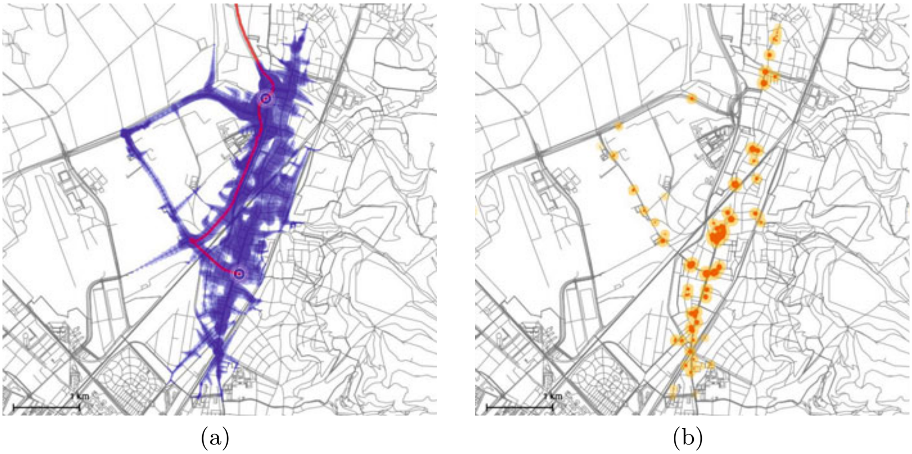


Fig. 4. (a) Reachable area between two published locations (4 min 56 sec). The radius of the circles indicate the possible waiting time to reach the final goal in time. The red line shows the actual route the user took. (b) The marked areas indicate possible visited POIs; the radius indicates the possible length of stay.

[2] http://wiki.openstreetmap.org/wiki/Map_Features#Landuse, [1/15/2011].

The dynamic measurement can be refined by incorporating all possible reachable plausible positions within a given timespan and velocity. Another possibility is to include all reachable POIs and the maximum possible length of stay. Figure 4(a) illustrates a possible implementation of *numreach*(), i.e., calculating all reachable plausible positions between two consecutive location disclosures. In this example the assumption about the potential travel speed was held static. This restriction can be lifted by using a more sophisticated route-planning algorithm and further external information. Figure 4(b) shows possible reachable POIs and the possible duration of stay.

6 Conclusion and Outlook

As today's communication scenarios get more diverse, the assumption of the anonymity of a user when sharing location data seems inadequate in many cases. As location information gains in importance, every entity involved in the communication process has to be considered as an adversary, since communication peers are usually considered as partially trusted.

We have proposed a theoretical user-centric privacy metric to allow a user to uncover the extent of information disclosure and to evaluate autonomously his privacy level in a communication relation with semi-trusted listener groups. The model makes no assumptions about the adversary's knowledge, capabilities or intention. The goal of such a metric is to enable an individual to estimate the knowledge gain caused by disclosing further location information in a specific communication scenario and thus enabling the user to make informed choices, e.g., choose the right protection mechanism. We divided the location privacy level into different subcomponents, reflecting the user's trust level, periodic habits leading to re-identification or to uncovering personal preferences, the evaluation of the user's exposure at a given time and differentiated between different kinds of location samples. Finally, some examples of an implementation for the main components were discussed.

For future work an implementation in a real-world application has to be developed together with a simple visualization of the privacy metric result. Therewith (location) privacy becomes for the user a more concrete fact instead of simply an abstract definition.

References

1. Westin, A.F.: Privacy and Freedom, 1st edn., Atheneum, New York (1967)
2. Duckham, M., Kulik, L.: Location privacy and location-aware computing, pp. 35–51. CRC Press, Boca Rator (2006)
3. 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(8), 1089–1107 (2010)
4. 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)

5. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5), 557–570 (2002)
6. 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)
7. Gedik, B., Liu, L.: Protecting location privacy with personalized k-anonymity: Architecture and algorithms. IEEE Transactions on Mobile Computing 7(1), 1–18 (2008)
8. Liu, L.: Privacy and location anonymization in location-based services. SIGSPATIAL Special 1, 15–22 (2009)
9. 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)
10. Beresford, A., Stajano, F.: Location privacy in pervasive computing. IEEE Pervasive Computing 2(1), 46–55 (2003)
11. Cranshaw, J., Toch, E., Hong, J., Kittur, A., Sadeh, N.: Bridging the gap between physical location and online social networks. In: Ubicomp 2010: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, pp. 119–128. ACM, New York (2010)
12. Toch, E., Cranshaw, J., Drielsma, P.H., Tsai, J.Y., Kelley, P.G., Springfield, J., Cranor, L., Hong, J., Sadeh, N.: Empirical models of privacy in location sharing. In: Ubicomp 2010: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, pp. 129–138. ACM, New York (2010)
13. 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)
14. Kamiyama, K., Ngoc, T.H., Echizen, I., Yoshiura, H.: Measuring Accumulated Revelations of Private Information by Multiple Media. In: Cellary, W., Estevez, E. (eds.) Software Services for e-World. IFIP AICT, vol. 341, pp. 70–80. Springer, Heidelberg (2010)
15. Golle, P., Partridge, K.: On the Anonymity of Home/Work Location Pairs. In: Tokuda, H., Beigl, M., Friday, A., Brush, A.J.B., Tobe, Y. (eds.) Pervasive 2009. LNCS, vol. 5538, pp. 390–397. Springer, Heidelberg (2009)
16. 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)
17. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. Nature 453(7196), 779–782 (2008)
18. 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 2009, pp. 1–9 (2009)
19. Ashbrook, D., Starner, T.: Using gps to learn significant locations and predict movement across multiple users. Personal Ubiquitous Comput. 7, 275–286 (2003)
20. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (1951)