On the Privacy Implications of Location Semantics

Berker Ağir, Kévin Huguenin, Urs Hengartner, Jean-Pierre Hubaux

On the Privacy Implications of Location Semantics

This work contains preliminary results; experiments with more users in progress.

Berker Ağır, EPFL, Switzerland, berker.agir@epfl.ch
Kévin Huguenin, LAAS-CNRS, France, kevin.huguenin@laas.fr
Reza Shokri, UT Austin, the USA, shokri@cs.utexas.edu
Urs Hengartner, University of Waterloo, Canada, urs.hengartner@uwaterloo.ca
Jean-Pierre Hubaux, EPFL, Switzerland, jean-pierre.hubaux@epfl.ch

***Abstract*—Mobile users increasingly make use of location-based online services enabled by localization systems (e.g., GPS). Not only do they share their locations to obtain contextual services in return (e.g., 'nearest restaurant'), but they also share information about the venues (e.g., the *type*, such as a restaurant or a cinema) they visit with their friends. This introduces an additional dimension to the threat to location privacy: location semantics that, combined with location information, can be used to improve location inference by learning and exploiting patterns at the semantic level (e.g., people go to cinemas after going to restaurants). Conversely, the type of venue a user visits can be inferred, hence this knowledge can be used to aggravate the threat to her (semantic) location privacy. In this paper, we formalize this problem and analyze the effect of venue-type information on location privacy. We introduce two multidimensional inference models that consider location semantics and semantic privacy-protection mechanisms. We evaluate our models by using a real data-set of semantic check-ins from Foursquare (obtained through Twitter). Our experimental results show that users' semantic location privacy is at serious risk and that semantic information significantly improves inference of user locations, hence multidimensionally degrading location privacy.**

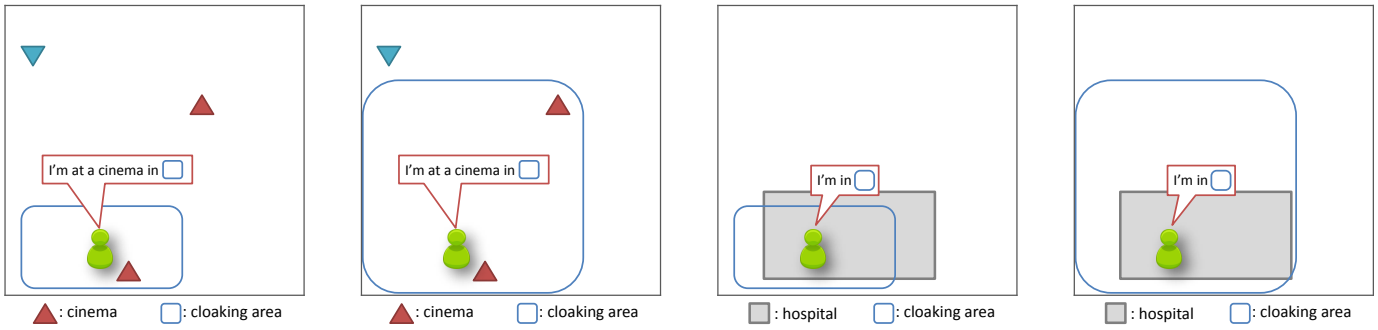
I. Introduction

Advanced localization-technologies and continuous Internet connectivity on mobile devices enable people to adopt an online life style; increasingly more people use mobile devices to receive location-based services and to enjoy location-based social platforms. Users of such systems provide location information (and possibly their identity) to the service providers in return for useful information such as the location of the nearest restaurant, cinema or nearby friends. Many of these services and systems are presented as free, but in fact, they obtain fine-grained user traces that can be used to infer more personal information: the price a user pays for benefitting from such services is her location data, which is detrimental to her privacy. This problem has been extensively investigated by the research community, the focus is mostly on geographical location privacy and related protection mechanisms [1]. Researchers have also studied how an adversary can locate/track users' whereabouts based on location samples that are, in some cases, anonymized and/or obfuscated, and on mobility history (*e.g.,* [2], [3]).

Many online service providers interact with their users on a multidimensional scale. For example, they let users state what type of location they are at, whom they are with and even what they feel at that specific time. This kind of data disclosure enables the service providers to enhance their knowledge about their users. Hence, the approach to location privacy from a purely geographical perspective is not sufficient anymore. Additional dimensions of information about the locations of users can be exploited by service providers, thus rendering privacy-protection mechanisms ineffective. Figure 1 depicts two examples where the semantic dimension (*i.e.,* the type) of location can be exploited to infer the actual location and the semantics of the user's location is not being protected at all. In Figure 1a, we observe that a user who visits a cinema discloses that she is in the depicted cloaking area and at a cinema. The problem is that there is only one cinema in this cloaking area, hence any observer of this disclosure can pinpoint the user easily. In another example, in Figure 1c, a user is at a hospital and wants to protect her location privacy. Unfortunately, her cloaking area is mostly occupied by the hospital, hence even though her exact location might not be pinpointed, the fact that she is at a hospital can be inferred without much confusion.

In this paper, we consider the case where users disclose not only their (obfuscated) geographic locations but also the types of venue they are at (*e.g.,* restaurant) in the form of check-ins (*e.g.,* social networks). Note that, unlike the approaches proposed in [4], [5], [6], [7], the venues visited by the users are assumed to be reported directly by the users and not *inferred* from the users' locations. We focus on the semantic dimension of the location and study its effects on location privacy, both on the geographical and semantic levels. Even though previously, researchers proposed semantic-aware protection mechanisms that aim to protect both geographical and semantic location privacy (*e.g.,* [8], [9]), their attempts lacked the extensive evaluation of their methods against a concrete adversary model. We fill this gap by proposing a model of an adversary; it considers a semantic-driven user-mobility behavior to multidimensionally attack location traces. We evaluate both geographical and semantic location-privacy with this adversary model and show that disclosing information about the type of location, *i.e.,* semantic location-information, decreases geographical location-privacy as much as 60%. We also present the threat on semantic location-privacy that deteriorates as quickly as the adversary gains background information on user-mobility profiles, that are easier to crawl by using publicly available data on various social networks. To the best of our knowledge, this is the first work that quantifies semantic location-privacy and demonstrates the effects of location semantics on location privacy.

The remainder of the paper is organized as follows: We



(a) (b) (c) (d)

Fig. 1: Examples showing how location semantics can be exploited when attacking location privacy and how semantic location-privacy is at risk. (a) A user reports that she is in the depicted cloaking area and also that she is at a cinema. Clearly, her location can be pinpointed by anyone who has access to this report, because there is only one cinema in the user's reported cloaking area, and this cinema occupies a small area compared to the cloaking area. (b) demonstrates how the ineffectiveness in (a) can be reduced by enlarging the cloaking area to include another cinema. An adversary can still trim the cloaking area, but now there are two locations with the tag cinema. (c) A user at a hospital now reports a cloaking area without revealing her semantic information. Clearly, this user does not want to disclose that she is at a hospital, but the hospital occupies a very big part of the cloaking area making it easy for an adversary to infer that she is at a hospital. These kinds of ineffective protection mechanisms put the semantic location-privacy of users at great risk. (d) demonstrates how semantic location privacy can be protected better by generating large cloaking areas to avoid domination of only one type of location in the reported cloaking areas.

introduce the reader to the context and define the system model in Section II. In Section III, we present our adversary model, inference approach and describe how we measure privacy. We explain our experimental setup and the datasets in Section IV and report the evaluation results. In Section V, we survey related work. Finally, in Section VI, we conclude the paper and discuss future work.

II. Background and System Model

We consider mobile users equipped with smartphones that have GPS modules and Internet connectivity. These users move in a geographical area and make use of location-based online services and social networks. We consider that users sporadically report their (potentially obfuscated) locations and, in some cases, semantic information (*i.e.,* the type, in the form of tags such as 'restaurant') of their locations. Note that we do not consider the case where the adversary extracts semantic information from the users' location traces, as considered in *e.g.,* [4]. In this setting, we consider an honest-but-curious service provider that is interested in inferring, based on its observations, users' actual geographical locations and the semantic tags associated with them, if any. Table I lists the notations used throughout the paper.

A. Users

Mobile users with GPS-equipped devices move in a given geographical area that is partitioned into M geographical regions $\mathcal{R} = \{R_1, R_2, \dots R_M\}$. Geographical areas are usually coarse-grained (typically cells associated with cell towers or regular square tiles of a several hundreds of meters). A subset of or all the areas in $\mathcal{R}$ contain venues annotated with semantic tags from the set $\{S_1, S_2, ..., S_K\}$, *i.e.,* a predefined list of keywords (*e.g.,* Foursquare defines such a list and all registered venues are tagged with such a keyword). Whenever a venue is visited by a user, it is mapped to the geographical region from $\mathcal{R}$ it falls in. We denote by $\perp$ the semantics of regions for the case when a user is in a geographical region, but does not visit a particular venue with a semantic tag, meaning that her location does not have semantic information. Hence, we define the set $\mathcal{S}$ of semantic tags as the union $\{S_1, S_2, ..., S_K\} \cup \{\perp\}$ to cover all semantic cases. Moreover, we consider discrete-time and a limited-time period $\{1, \dots, T\}$.

As users move, they sporadically use online services and share their (potentially obfuscated) locations together with the corresponding (potentially obfuscated) semantic tags. Formally, whenever a user u visits a geographical region r at a time instant $t \in \{1, ..., T\}$, she generates an event consisting of her actual geographical region $r \in \mathcal{R}$ and the corresponding semantic tag $s \in \mathcal{S}$. This user event at time instant t is denoted by $a_u(t) = (r, s)$; in other words, the *actual location* of user u at time instant t is represented by the pair (r, s). We denote by $a_u = \{a_u(1), ..., a_u(T)\}$ the whole trace of user u.

B. Privacy Protection Mechanisms

For privacy reasons, users employ privacy-protection mechanisms (PPMs) before reporting their location and semantic information to an online service provider[2]. Typically, a PPM,

[1] $\mathcal{P}$: Power set.

[2] We refer to the *online service provider* as the *service provider* or the *adversary* for short in the remainder of the paper.

TABLE I: Table of Notations

$\mathcal{R}$	Set of geographical regions
$\mathcal{S}$	Set of semantic tags
$a_u(t) = (r, s)$	User u's actual event at time instant t, where $r \in \mathcal{R}$ and $s \in \mathcal{S}$
$o_u(t) = (r', s')$	User u's obfuscated event at time instant t, where $r' \in \mathcal{P}(\mathcal{R})$[1]
a_u	Actual trace of user u
o_u	Obfuscated trace of user u
$\mathbf{R}_t, \mathbf{R}'_t$	The actual and obfuscated location variables for time t
$\mathbf{S}_t, \mathbf{S}'_t$	The actual and obfuscated semantic variables for time t
$f_u(r, r')$	A geographical PPM modeled as a probability distribution function employed by user u
$g_u(s, s')$	A semantic PPM modeled as a probability distribution function employed by user u

that aims to protect the geographical location of a user, replaces her actual location with another location (*i.e.,* perturbs the location) or with a list of locations (*i.e.,* a cloaking area), or hides the location information completely. In this work, we consider such PPMs and the PPMs that protect the semantic dimension of the location, specifically the semantic tag of a user's event. In particular, these PPMs generalize the semantic tag (*i.e.,* introduce additional or more generic tags) or hide it completely. We assume that a set of PPMs obfuscates a user's actual event at time t independently from her other events at other time instants. Such a PPM model can also cover the cases where the underlying localization technique used by the adversary returns coarse-grained and possibly bogus information about the users.

After applying PPMs on her actual geographical region r and the corresponding actual semantic tag s, a user u reports her obfuscated geographical region r' and the obfuscated semantic tag s' to the service provider. r' (resp. s') is typically a subset of $\mathcal{R}$ (resp. $\mathcal{S}$). We assume that the service provider only observes the obfuscated trace $o_u = \{o_u(t) = (r', s')\}, \forall t \in \{1, 2, ..., T\}$ of user u.

A PPM is successful to the extent that it confuses the adversary and leads him to believe that the imprecise parts of an obfuscated event are the user's actual event. In this sense, we model a PPM as a probability distribution function that maps actual events to obfuscated ones. Specifically, we denote by functions $f(r, r')$ and, respectively, $g(s, s')$ the probabilities to generate the obfuscated location r' and the obfuscated semantic tag s' (*i.e.,* $\mathbf{Pr}(r'|r)$ and $\mathbf{Pr}(s'|s)$) that constitute the obfuscated event $o(t) = (r', s')$ given the actual event $a_u(t) = (r, s)$. Note that geographical and semantic information are obfuscated independently from each other at each time instant and independently from the other time instants. Finally, we assume that users do not collaborate to protect each others' privacy and their identities are not anonymized.

C. Adversary

The adversary is typically a service provider or an external observer who has access to obfuscated traces of users. He has two main purposes: (*i*) locate users at specific time instants, and (*ii*) identify the types of the locations users visited in terms of the semantic tags associated with them. While carrying out his attack, the adversary takes into account the relationship between geographical and semantic dimensions of location, which is explained in Section III.

The adversary runs his attack *a posteriori*, *i.e.,* after having observed the whole obfuscated trace o_u of a user u. Even though the obfuscation of an event is done independently from the other events of the user, the adversary assumes that a user's actual events are correlated and therefore models user mobility-behavior. He is assumed to have access to users' (partial) past events that, he exploits to build a mobility profile for each user u, consisting of both geographical and semantic dimensions. Basically, a user's mobility profile represents the user's transition probabilities regarding mobility, *i.e.,* between geographical regions and between semantic tags. Formally, such a mobility profile is the set of the probability distribution functions $\mathbf{Pr}(r|\rho)$, $\mathbf{Pr}(s|\sigma)$ and $\mathbf{Pr}(r|s)$, where ρ and σ represent the user's previous location and previous semantic tag.

The adversary also knows which PPMs a user u employs and with what parameter(s), *i.e.,* the functions f_u and g_u. Together with the PPMs and the mobility profile he generates, the adversary performs his attack on a user trace given her obfuscated trace o_u.

III. Inference and Privacy

We explain our model of inference and background knowledge of the adversary in the subsequent subsection. In summary, we build two user behavior models by using Bayesian networks under the assumption that people follow a Markovian mobility process. These models take into account both the geographical and semantic dimensions of the location and also the relationship between them. Based on these two models, we evaluate geographical and semantic location privacy.

A. Inference and Background Knowledge

We model the adversary's inference considering a specific way of user behavior based on a simple idea: users move based on what they plan to do next given their current context, *i.e.,* in this case, their locations and semantic information. We determine the following two scenarios (illustrated in Figure 2):

1) The adversary knows the users' geographical transition histories, *i.e.,* the *geographical background*, and assumes that the users move to new locations primarily based on their current locations. Their semantic tags are merely dependent on their locations. For instance, a user might go to a location in downtown after visiting another location in nearby downtown. The semantics of these locations then, for instance, might happen to be a cinema and a restaurant.

2) The adversary knows both the users' geographical transition histories and semantic transition histories, together referred to as *geographical & semantic background*. Unlike the first scenario, in this case the user first determines what type of place she will go to (*i.e.,* the next semantic tag) given the semantic tag of her current location, and then chooses the region she will go to based on the determined next semantic tag and her current location. For instance, if a user is at a restaurant in downtown and wants to go to a cinema, she chooses to go to a cinema that is close to her current location (that she often visits).

We elaborate more on these scenarios and their respective Bayesian networks in the following sections.

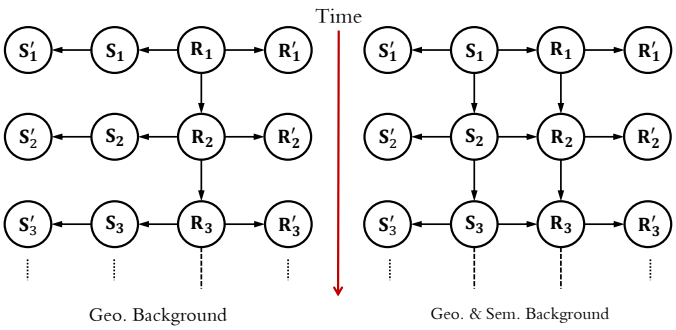


Fig. 2: The Bayesian networks representing our adversary's user modelling. The one on the left prioritizes the geographical transitions with only geographical background known to the adversary. The right one prioritizes the semantic transitions over geographical transitions with both geographical and semantic background. Protection mechanisms work separately on regions and semantic tags and they are independent from each other.

1) Geographical-Only Background: As stated previously, the adversary knows the users' geographical transition histories in this scenario and wants to carry out his attack based only on this type of available information. He can correlate the sequential events of a user by using only geographical background information, hence we build a Bayesian network in which only the region (*i.e.,* the geographical location) nodes are connected to each other among user events. But, as the adversary still wants to infer the semantic tags in the user events, semantic nodes are also created and they are dependent on the region nodes. This ensures that the adversary benefits from any semantic information disclosed by the users in his inference, even though he does not have any semantic background information. In summary, with this model, a users is assumed to determine her next location (and indirectly the next semantic tag) based on her current location.

This model is illustrated in Figure 2 on the left, where each line of nodes represent a user event in time, both actual $(\mathbf{R}_t, \mathbf{S}_t)$ and obfuscated $(\mathbf{R}'_t, \mathbf{S}'_t)$, where $\mathbf{R}_t$, $\mathbf{S}_t$, $\mathbf{R}'_t$ and $\mathbf{S}'_t$ are the network node names and random variables for a user's actual and obfuscated events at time t. The conditional probability distributions for the obfuscated events', *i.e.,* for $\mathbf{R}'_t$ and $\mathbf{S}'_t$), are the privacy-protection mechanism distributions f_u and g_u, explained in Section II-B. If a static privacy-protection mechanism (PPM) is used by the users, then these functions map the actual regions and the actual semantic tags to obfuscated regions and obfuscated semantic tags with probability 1 (*i.e.,* for a given region, resp. a semantic tag, the PPM always generates the same obfuscation outcome). More confusing PPMs can be employed and used in this network, *e.g.,* hiding the actual information completely with a hiding probability.

The remaining conditional probabilities are the user's actual semantic tag given her actual location $\mathbf{Pr}(\mathbf{S}|\mathbf{R})$ and the user's next location given her current location $\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t)$. We calculate $\mathbf{Pr}(\mathbf{S}|\mathbf{R})$ based on the semantic tags' associations to regions as the adversary is assumed to have no semantic background information. Basically, $\mathbf{Pr}(\mathbf{S}|\mathbf{R})$ represents a uniform distribution over all semantic tags associated with a region r, *e.g.,* if a region has 4 semantic tags (including the $\perp$ tag) associated with it, then the probability for each of these tags to be the actual tag given this location is 0.25. Lastly, we compute $\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t)$ by counting the number of transitions among all regions in a user trace and then using the Markovian knowledge construction approach used in [3]. For the root node in the network, which is $\mathbf{R}$ for the geographical background scenario, we use the steady-state probability distribution computed from $\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t)$.

2) Geographical and Semantic Background: In this scenario, we consider an adversary that models user mobility-behavior in an activity-driven fashion: A user first determines the type (*i.e.,* the semantic tag) of her next geographical region given the type of her current geographical region; then, she determines the next geographical region given her current geographical region *and* the next semantic tag. For example, a user decides to go to a restaurant, then she chooses which restaurant she wants to go. Afterwards, she wants to go to a cinema. Considering her previous location, she picks the cinema that is most convenient for her. This model is depicted in Fig. 2 on the right-hand side.

As in the scenario with only the geographical background knowledge, the conditional probability distributions for the obfuscated events (*i.e.,* $\mathbf{R}'_t$ and $\mathbf{S}'_t$) are the same. Whereas, the transitions between user events now require a semantic-transition distribution ($\mathbf{Pr}(\mathbf{S}_{t+1}|\mathbf{S}_t)$) and a geographical-transition distribution, which is also conditioned on the semantics of the next user-event ($\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t, \mathbf{S}_{t+1})$), meaning that $\mathbf{R}_{t+1}$ depends on the user's current semantic tag $\mathbf{S}_{t+1}$ and her previous geographical region $\mathbf{R}_t$.

The semantic transition distribution $\mathbf{Pr}(\mathbf{S}_{t+1}|\mathbf{S}_t)$ is constructed in the same way the geographical transition distribution $\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t)$ is constructed. However, as we consider that geographical and semantic background information separately, the adversary is assumed not to know the distribution $\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t, \mathbf{S}_{t+1})$. In short, the adversary is assumed to have knowledge on $\mathbf{Pr}(\mathbf{S}_{t+1}|\mathbf{S}_t)$, $\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t)$ and $\mathbf{Pr}(\mathbf{R}_t|\mathbf{S}_t)$ to some extent regarding user history. There-

fore, he needs to use $\mathbf{Pr}(\mathbf{R}_t|\mathbf{S}_t)$ and $\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t)$ to derive $\mathbf{Pr}(\mathbf{R}_{t+1}|\mathbf{R}_t, \mathbf{S}_{t+1})$. We achieve this simply by normalizing the marginal probability distribution $\mathbf{Pr}\{\mathbf{R}_{t+1}|\mathbf{R}_t\}$ for a given semantic tag s (*i.e.,* over regions that have s) and by combining it with the conditional distribution $\mathbf{Pr}(\mathbf{R}_t|\mathbf{S}_t = s)$. For the rest of the geographical regions, *i.e.,* those that do not have the semantic tag s, the probability is simply 0. This translates to the following formula:

$$\mathbf{Pr}(\mathbf{R}_{t+1} = r|\mathbf{R}_t = \rho, \mathbf{S}_{t+1} = s) = \tag{1}$$
$$\begin{cases} 0 & \text{if } s \notin r \\ \left(\alpha \cdot \dfrac{\mathbf{Pr}(\mathbf{R}_{t+1} = r|\mathbf{R}_t = \rho)}{\sum\limits_{R_m, s \in R_m} \mathbf{Pr}(\mathbf{R}_{t+1} = R_m|\mathbf{R}_t = \rho)} + (1-\alpha) \cdot \mathbf{Pr}(\mathbf{R}_{t+1} = r|\mathbf{S}_{t+1} = s) \right) & \text{otherwise} \end{cases}$$

, where α is the factor to set the weight of geographical transitions against the probability that $\mathbf{R}_{t+1}$ is r given $\mathbf{S}_{t+1} = s$ (which is derived from the number of visits to a region r given the semantic tag s in the user history). In other words, α is used to control how much importance is distributed among different types of user history, *i.e.,* geographical transitions and steady user events. In our experiments, we set α to 0.5, which we believe is a balanced treatment of user history. Note that this separate treatment of the geographical and semantic background information enables the adversary to exploit the semantic mobility of a user's behavior data in one city to infer user events in another city, where he might lack the knowledge.

Note that the aforementioned models might not reflect the users' actual behaviors. However, such models (in particular the Markovian mobility assumption) are widely used in practice (and considered in the literature) as they enable the adversary to develop efficient algorithmic and computational methods to infer the users' locations. The accuracy of the inference attack carried out by the adversary partially depends on how well the user model fits the users' actual behaviors.

B. Privacy Measurement

Due to different privacy concerns in both geographical and semantic dimensions of location, we measure the privacy level in both dimensions separately. Privacy levels in both dimensions are measured as a function of the expected error of the adversary. The inference based on our Bayesian networks yields probability distributions over regions and semantics that fit this measurement approach. In other words, the output of the inference is a PDF for each node in a given Bayesian network, *i.e.,* the PDF h_g over all regions at every time instant for user location and the PDF h_s over all semantic tags at every time instant for user semantic tag. The geographical and semantic privacy levels of a user u at time instant t, denoted by $\text{GP}_u(t)$ and $\text{SP}_u(t)$, are computed as follows:

$$\text{GP}_u(t) = \sum_{m=1}^{M} h_g(R_m, t) \cdot \text{dist}^G(R_m, r), \tag{2}$$

$$\text{SP}_u(t) = \sum_{k=1}^{K} h_s(S_k, t) \cdot \text{dist}^S(S_k, s), \tag{3}$$

where $\text{dist}^G(\cdot,\cdot)$ and $\text{dist}^S(\cdot,\cdot)$ are geographical and semantic distance functions, and (r, s) is the actual event of user u at time instant t.

We use the Haversine formula[3] to compute the geographical distances between regions by using the coordinates of their center points; we use the binary distance for the semantic distances, meaning that if two semantic tags are the same, then then distance is 0, otherwise 1.

IV. Evaluation

We evaluate both geographical and semantic location-privacy in a joint way. More specifically, we consider user events that have both geographical and semantic location information. We analyze privacy by running experiments on a real dataset that not only has geographical location data but also semantic information in most cases (see Section IV-A). In our experiments, we study the effects of semantics on the geographical location-privacy compared to semantic-oblivious studies. More specifically, we would like to find out if semantics help an adversary infer more accurately the actual geographical locations of users, hence whether they can potentially be used to design more powerful PPMs compared to the existing work. We also analyze the effect of geographical information on semantic location-privacy from similar aspects.

A. Dataset

In order to experimentally evaluate users' semantic location-privacy and the effect of semantic information on users' location-privacy, we rely on a dataset of real user check-ins, which include geographic and semantic information about the venues visited by the users of a large location-based social network (we crawled it). In addition, we rely on a predictive utility model based on user feedback collected through a personalized online survey targeted at Foursquare users ($N = 77$) recruited via Amazon Mechanical Turk (obtained from [10]). In this section, we give details about our data sources, including the data collection, filtering and processing methodology and general descriptive statistics about the data.

1) Location Traces with Semantics: As a starting point, we use a tweet dataset we collected between January 2015 and July 2015. The dataset contains public geo-tagged tweets (i.e., Twitter lets users to attach their GPS coordinates to their tweets) from all over the world with a focus on six cities (New York, San Francisco, Boston, Chicago, London and Istanbul). We collected these tweets by identifying users through Twitter's public tweet stream (i.e., $\sim$1% of the Twitter public timeline) and by fetching timelines of these users. The dataset contains a total of $\sim$72 million geo-tagged tweets generated by 1,493,287 unique users. A summary of the statistics of the dataset is provided Table II.

The coordinates embedded in the geo-tagged tweets, however, do not contain semantic information (which we need

[3]This formula is used to compute the distance between two points on a given sphere, in our case the Earth.

[4]We present preliminary results with this initial set of users. We are running experiments with more users.

TABLE II: Dataset Statistics

	raw dataset	filtered dataset
# unique users	1,493,287	10^4
# geo-tagged tweets	∼72 mil.	1214
# FS check-ins	∼14.3 mil.	1124
# Distinct FS tags	649	158

TABLE III: Experimental Setup

Number of iterations	20
Area size	1.6×1.6 km^2 (8×8 regions)
Average Proportion of FS Tweets per user (*i.e.,* tweets w/ semantic information)	92%

for our evaluation). To obtain such information, we rely on Foursquare. Foursquare is a large location-based social network that enables its users to check-in at nearby venues (selected from the Foursquare database of registered and confirmed venues). It offers its users the option of linking their Foursquare accounts with their Twitter accounts in such a way that, whenever a user checks-in, Foursquare generates a related text with a short URL to the Foursquare check-in and tweets it, along with the GPS coordinates, on the user's Twitter timeline. We select such Foursquare-generated tweets from our Twitter dataset and, for each, we parse the URL to the Foursquare check-in from the tweet text. Using these URLs, we fetch (through the Foursquare API) the corresponding check-in and the venue. For each venue referenced in a check-in of our dataset, we collect rich statistical information such as total number of visits, total unique visitors, rating, etc. Most importantly, we collect the coordinates[5] and the semantic tag(s) (a primary tag and possibly a secondary tag), selected from a pre-defined set of 763 tags (*i.e.,* referred to as Foursquare categories) organized as a tree, assigned to the venue. We identify 649 distinct tags assigned to the venues in our dataset. Finally, we demonstrate the correlation between the venue density and the FS tweet density in geographical places in Figure 4, which shows a venue heat map and a Foursquare check-in heat map in San Francisco Bay Area.

In our evaluation, we use a subset of the dataset (due to computational limitations): We focus on the check-ins made in a geographical region in the San Francisco Bay area (i.e, (37.77504, -122.406775) to (37.7894, -122.42496), of size approximately 1.6×1.6 km^2) and extract users with at least 70 tweets in this region. We further filter out users whose FS tweets (*i.e.,* check-ins) account for less than 70% of all their tweets (*i.e.,* most of the tweets user in the experiments contain venue information). The final dataset results in 10 users[6]; see Figure 5 for users' count of FS and other tweets. The maximum number of tweets per user observed in the filtered dataset is 176. We included all the tweets of a user

[5]Note that GPS coordinates in the tweets might differ from registered venue coordinates at Foursquare due to inaccuracy of GPS modules on mobile devices. In such cases, we use the coordinates of the venues.

[6]We are currently working on collecting a dataset with more users.

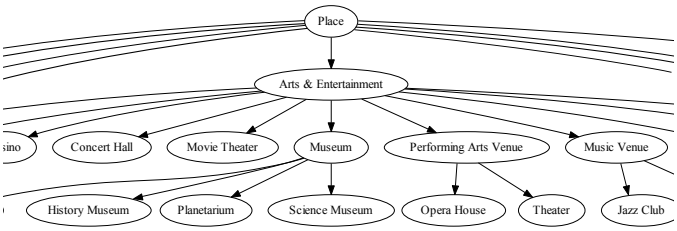
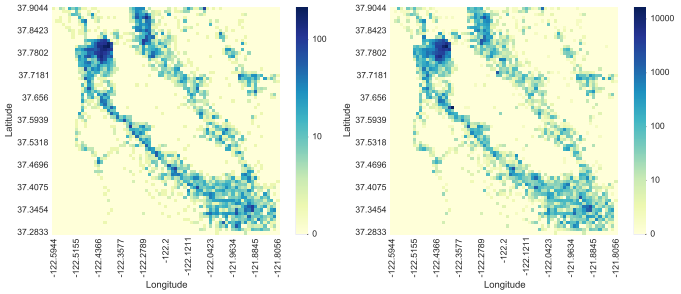


Fig. 3: Part of the Foursquare category hierarchy that we use as our semantic tag tree. 'Place' tag is the root.



(a) Venue heat map (b) FS Check-in heat map

Fig. 4: Venue and check-in heat maps (*i.e.,* count distribution) in greater San Francisco Bay Area.

in the knowledge construction of the adversary and for each user we use a randomly selected subtrace of length 5 in each experiment. There are 1341 venues in our filtered dataset and the tag distribution over these venues has a heavy tail as shown in Figure 6.

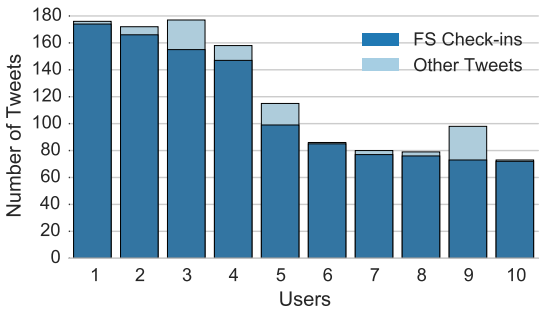


Fig. 5: Number of FS check-ins/tweets and other tweets per user in the filtered dataset used in our experiments.

2) Predictive Utility Model: Semantic obfuscation, usually achieved through generalization as discussed in the previous sections, is likely to have a negative effect on the utility of the service as perceived by the users. As the notion of (perceived) utility is quite subjective, user feedback is needed to model and quantify the utility implications of the use of obfuscation techniques. In order to build such a model, we rely on a data-set collected by the authors of [10]. In this work, the authors performed a personalized survey with 77 active Foursquare users recruited through Amazon Mechanical Turk. In the survey, each participant was shown 45 of her own past Foursquare check-ins; for each of these check-ins,

TABLE IV: Example of obfuscated check-ins with different combinations of geographical and semantic obfuscation (source: [10]).

Obfuscation levels	**Example**
Original check-in	The Westin Hotel, 320 N Dearborn St. (Chicago 60654, IL, United States)
Low semantic, Low geographical (Ls-Lg)	At a hotel, on Dearborn St. (Chicago 60654, IL, United States)
High semantic, Low geographical (Hs-Lg)	At a travel & transport place, on Dearborn St. (Chicago 60654, IL, United States)
Low semantic, High geographical (Ls-Hg)	At a hotel, in Chicago (IL, United States)
High semantic, High geographical (Hs-Hg)	At a travel & transport place, in Chicago (IL, United States)

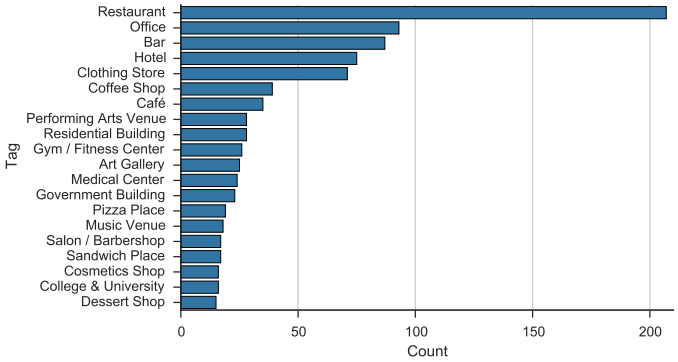


Fig. 6: Number of venues per semantic tag in the filtered dataset for the top 20 tags.

the participant was presented with four different obfuscated versions of the check-in and she was requested to rate, on a scale from 1 to 5 (where 1 is "not at all" and 5 is "perfectly"), to what extent the purpose of her check-in would still be met if the precise venue location was replaced with the obfuscated version of it. The four obfuscated versions of the check-in were generated by applying the possible combinations of low/high semantic obfuscation (Ls or Hs) and low/high geographical obfuscation (Lg or Hg) as illustrated in Table IV (extracted from the original article). One finding from the article is that semantic obfuscation has a higher negative effect on utility than geographical obfuscation does.

Using this data, to predict the utility of an obfuscated version of a check-in (on a discrete scale from 1 to 5), the authors propose a utility model that relies on a number of features extracted from the users' check-in, including the check-in location, date, time, text, and the venue type. The predictive model proposed in the original paper achieves high accuracy with a median error of around 0.5. In order to quantify utility, we build a simplified version of the predictive utility model proposed in [10] (based on the same data). Our model is based on only two different features: the venue type and the obfuscation level. The median error of our simplified model is 1.1, which is sufficient for our purpose (*i.e.,* exploring the privacy-utility trade-off).

B. Experimental Setup

Methodology: We partitioned the considered area into 64 square regions using an 8×8 regular grid. Within this area, we identified 1341 Foursquare venues over 158 unique semantic tags. We then mapped the users' traces to this setting. We implemented our Bayesian network-based models on Python by using the Bayesian Belief Networks library provided by eBay [11]. We applied certain protection approaches on the users' mapped traces, set these *protected*/observed traces in our Bayesian networks as observations, and applied the junction-tree inference algorithm [12].

Privacy Measurement: We evaluate the privacy as the error of the adversary. The larger the error is, the better the privacy is. For this, we compute the expected error on the inferred locations. We use the Haversine distance when measuring the geographical location-privacy. For this, we use the geographical coordinates of the center points of the square regions in our setting. For semantic location-privacy, we use binary distance, *i.e.,* the distance between two distinct tags is always considered 1 and 0 for the same tags.

Background Knowledge: In our experiments, the adversary always has geographical background knowledge on the users' history (*i.e.,* transitions). Based on this we have two different scenarios (explained in detail in Section III-A):

1) **Geographical Background:** In this scenario, the adversary is assumed to have knowledge on geographical transition patterns of users and no semantic background information. We run experiments for this scenario by using our first Bayesian network model that prioritizes the geographical transitions for user behavior introduced in III-A. The transitions are built using the number of geographical transitions in the whole traces of users.
2) **Geographical and Semantic Background:** The adversary is assumed to know more about users' histories: transitions in both geographical and semantic dimensions. He also knows the distribution of geographical region visits, given the semantic information on user traces, *i.e.,* how many times a region r was visited, given that the user event's semantic tag was s. This type of background information enables us to use our second Bayesian network model that prioritizes the semantic transitions for event sequences, meaning that the users move by first choosing the semantic tag of the location they want to go to and then determine a geographical region associated with this semantic tag based on their previous location.

Protection Mechanisms: We implement geographical and semantic location-privacy protection approaches separately, meaning that geographical protection does not take into account the semantic information of the user's actual location, and vice versa.

We implement a geographical location-privacy protection mechanism as an obfuscation mechanism that either generates an obfuscation area of a certain size or hides the geographical location completely with a predetermined probability (called hiding probability λ). This mechanism replaces any given region (*i.e.,* the actual location of a user) with a larger, square area in our map. For instance, a 2×2 obfuscation: (*i*) with probability $1 - \lambda$, generates an obfuscation area consisting of 4 adjacent regions, one being the actual location of the user, or (*ii*) with probability λ, hides the location.

We consider the following four scenarios regarding the semantic protection and, to compare their effects, employ each of them in separate experiments:

1) No protection. In this case, we directly disclose the actual semantic tag all the time.
2) Parent-tag obfuscation. This is a generalization based on the semantic tag tree derived from FS categories. Given the actual semantic tag of the user, we determine its parent tag in the tree and disclose this tag as the semantic information of the user's current location. For example, when the actual user event has the tag 'Theater', then according to the FS categories (which is partially reflected in Fig. 3), the parent tag 'Performing Arts Venue' is disclosed replacing 'Theater'.
3) Parent-tag obfuscation with hiding probability λ. In this case, we disclose the parent tag of the user's location with probability $1 - \lambda$ or hide the semantic information completely with hiding probability λ.
4) Hide the semantic tag completely. In this case, the adversary never observes any kind of semantic information from the users.

In our experiments, we employ the geographical protection mechanism in combination with each of the semantic protection/disclosure scenarios with varying hiding probabilities.

C. Experimental Results

In this section, we analyze the experimental results with different protection mechanisms in various settings. Due to the small number of users, the results presented in this section cannot be generalized and should be interpreted cautiously; experiments with more users are in progress.

1) Effect of Semantic Information on Location Privacy: We first investigate the effect of adding semantic information to a user's check-in on her geographical location privacy. We consider four protection scenarios with low to high granularity of semantic information combined with fixed geographical obfuscation over gradual hiding probability λ. Specifically, given a geographical obfuscation parameter (*e.g.,* 2×2 obfuscation) and for each λ, we evaluate four different semantic protection approaches (explained in Section IV-B) that are employed together with the obfuscation mechanism.

We present the results in Figure 7, where the x-axis represents the hiding probability λ (used for geographical obfuscation and parent-tag semantic generalization) and the y-axis represents the geographical location privacy in kilometers. We plot the geographical location privacy aggregated over all users, all events and all iterations of simulations for each protection mechanism and hiding probability (λ) pair using box plots. These box plots show the 1st, 2nd, 3rd quartiles of the related data and the 98% confidence intervals. We plot the semantic protection approaches employed together with the geographical obfuscation from the lightest box to the darkest in the following order:

- Hiding semantic information (lowest granularity)
  (Geo. (obf 2×2, λ) | Sem. ($\perp$, λ),
- Parent-tag generalization with hiding probability λ
  (Geo. (obf 2×2, λ) | Sem. (parent, λ)
- Direct parent-tag generalization
  (Geo. (obf 2×2, λ) | Sem. (parent, 0)
- Actual semantic tag (highest granularity)
  (Geo. (obf 2×2, λ) | Sem. (actual, λ)

where 'Geo. (obf 2×2, λ)' corresponds to 2×2 geographical obfuscation with hiding probability λ. We employed 2×2 and 4×4 obfuscation parameters in our experiments and we present the results hereafter.

We observe that as we disclose more semantic information, along with the obfuscated geographical location (from left to right for each λ value), the median location privacy consistently decreases in all cases. Also, unsurprisingly, the privacy level increases as we increase the granularity of the location (*i.e.,* from 2×2 obfuscation in Figure 7a to 4×4 obfuscation in Figure 7b). Note that for $\lambda = 1.0$, the parent-tag generalization with hiding probability λ is exactly the same as hiding the semantic information completely and, similarly, it is exactly the same as the direct parent-tag generalization (*i.e.,* always disclosing the parent tag instead of the actual tag) for $\lambda = 0.0$. These can be observed in Figure 7.

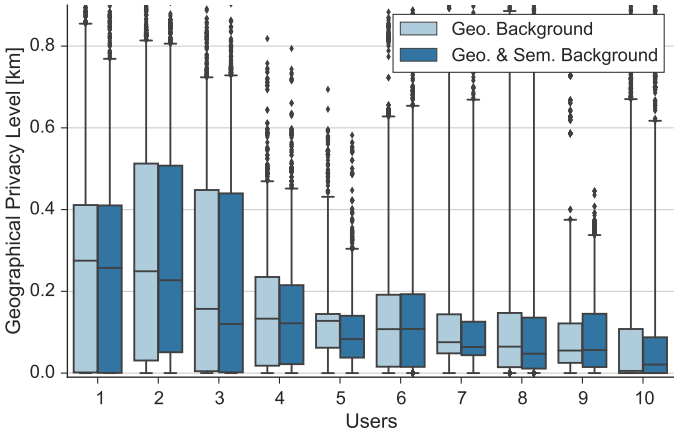
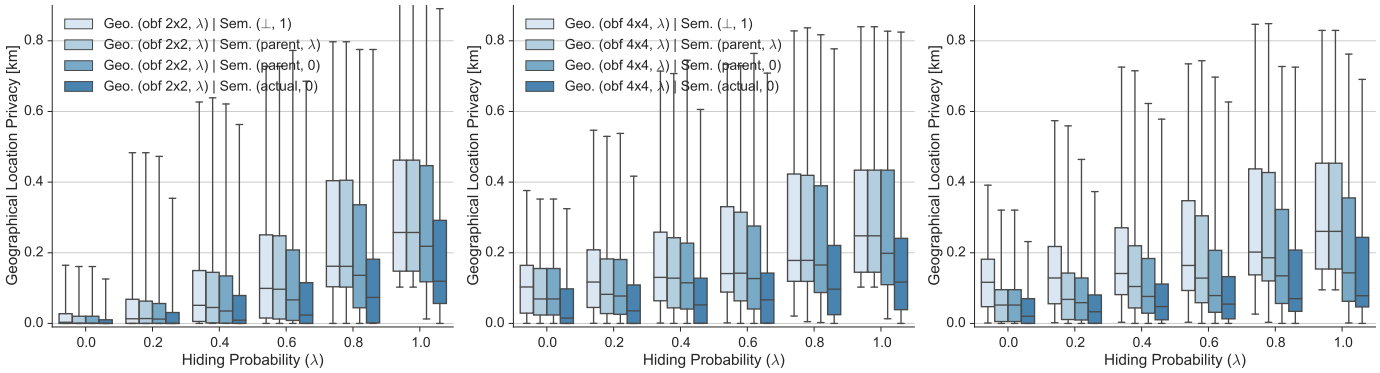


Fig. 8: Average geographical location privacy per user. The users are in descending order by their median privacy level.

We also analyze the effect of employing the semantic histories of users in inference, in addition to the geographical histories already employed in all our experiments (*i.e.,* semantic history in the form of semantic transitions such as people going to a *cinema* after going to a *restaurant*). We compare the two scenarios in case of the 4×4 geographical obfuscation with hiding probability λ (*i.e.,* figures 7b and 7c, with and without



(a) 2×2 Obfuscation w/o Sem. Background (b) 4×4 Obfuscation w/o Sem. Background (c) 4×4 Obfuscation w/ Sem. Background

Fig. 7: Geographical location privacy levels over different protection and learning scenarios.

semantic background information respectively). We observe that, for instance in case of $\lambda = 0.4$, the median geographical privacy decreases when the adversary employs the semantic background information of users. This pattern is visible for most of the cases from *without* semantic background to *with* semantic background. It is also visible that the semantic background information is very influential on geographical location privacy in the cases of direct parent-tag generalization and semantic disclosure (*i.e.,* the two darkest boxes). We notice that in some cases (typically for the light case where the semantic information is hidden all the time) the adversary is more confused (and hence less successful) when he employs semantic background knowledge. The main reason for this outcome is that the adversary's knowledge on the semantic transitions of the user is less effective in his attack when the attacked traces' length is short. In general, we observe that employing semantic background knowledge in the inference helps the adversary increase his accuracy from 10 to 60 meters. Figure 8 shows the effect of employing semantic background information, which is the general decreasing tendency in geographical location-privacy, in average for each user in an aggregated form (over all simulations, all λ values and all user events).

2) Privacy vs. Utility Trade-Off: We now explore the trade-off between privacy and utility by evaluating both (location) privacy and utility for different levels of obfuscation. We consider four protection mechanisms by combining a low or high level of semantic obfuscation with a low or high level of geographical obfuscation as described in Table V and illustrated in Figure 9. We set the hiding probability λ to 0.2.

TABLE V: Description of the different obfuscation levels.

Obfuscation	Description
Ls-Lg	Semantic tag, 2×2 geographical region
Hs-Lg	Parent semantic tag, 2×2 geographical region
Ls-Hg	Semantic tag, 4×4 geographical region
Hs-Hg	Parent semantic, 4×4 geographical region

We plot the results in Figure 10. The points represent the average privacy and utility. It can be observed that the four points corresponding to the different obfuscation levels form a diamond shape: Ls-Lg provides the highest level of utility and the lowest level of privacy; Hs-Hg provides the highest level of privacy but the lowest level of utility; Ls-Hg provides a better level of (location) privacy than Hs-Lg *and* a lower level of utility. This last observation is quite intuitive as geographical obfuscation is expected to protect location privacy better than semantic obfuscation and semantic obfuscation has been proved to be more detrimental to utility than geographical obfuscation has been [10]. This means that users should always prefer Ls-Hg over Hs-Lg.

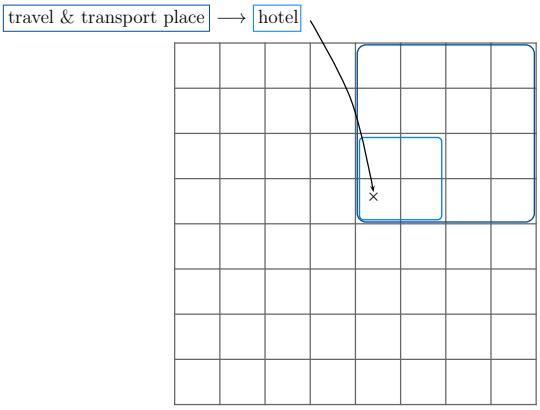


Fig. 9: Illustration of the obfuscation levels used in the experiments. Light blue frames denote low levels of obfuscations whereas dark blue frames denote high levels of obfuscation.

3) Semantic Location-Privacy: Finally, we evaluate the semantic location-privacy and present the loss of privacy in the semantic dimension of location. As with the geographical location-privacy figures, we plot the aggregated privacy-level over all users, all simulation iterations and all user events using box plots. The semantic location-privacy is calculated as the expected error of the adversary and, in this case, the error is binary: the distance (*i.e.,* dissimilarity) between two semantic tags is considered 1 if they are different and 0 if they are the same. Hence, the semantic privacy-level for any given user

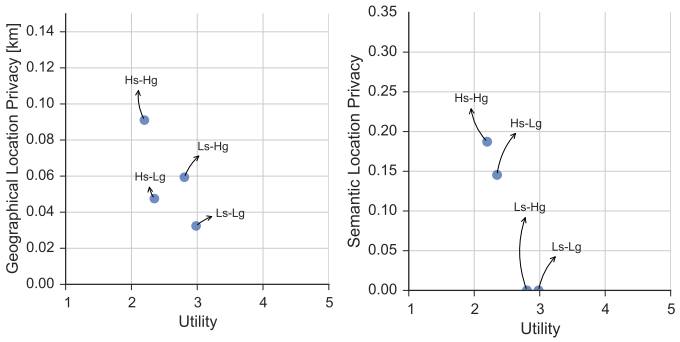
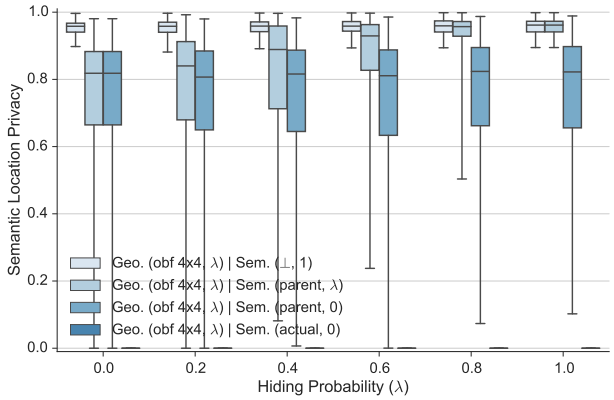
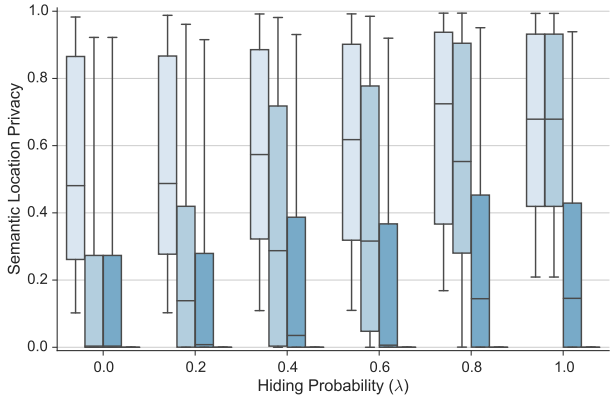


Fig. 10: Privacy vs. Utility



(a) 4×4 Obfuscation w/o Sem. Background



(b) 4×4 Obfuscation w/ Sem. Background

Fig. 11: Semantic location privacy levels over different protection scenarios with geographical and semantic background knowledge of the adversary.

event is always in the interval $[0, 1]$, 1 being the maximum level of privacy.

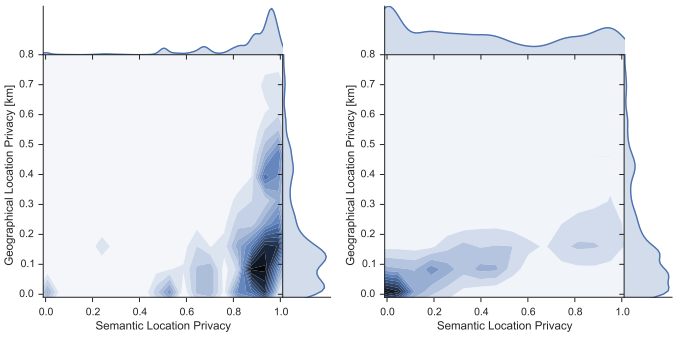
In Figure 11, we present the semantic location-privacy results for 4×4 obfuscation with hiding probability λ in both 'Geographical background' and 'Geographical & Semantic Background' scenarios. In both cases (shown separately in figures 11a and 11b), as we protect the semantic information of the users' traces less and less (from the lightest boxes to the darkest ones), the semantic location-privacy consistently decreases. We also observe that protecting the geographical location-privacy more, *i.e.,* increasing the hiding probability λ, also helps increase the semantic location-privacy in most of the cases. Whereas, semantic location-privacy is naturally always 0 in the case of disclosing semantic information all the time. Moreover, unsurprisingly, when the adversary has semantic background information in addition to the geographical one, he learns more about the users' location semantics in his inference, *i.e.,* the semantic location-privacy decreases. However, compared to the geographical dimension, this decrease in the semantic location-privacy is more significant as can be seen in Figures 11a and 11b: even if the semantic tags of the user events are hidden all the time, the privacy loss is between 30-50%. The loss reaches up to as much as 80% in other protection scenarios.

Lastly, we present the geographical and semantic location-privacy jointly in Figures 12a and 12b, without and with semantic background information, respectively. These plots represent the density of the privacy data over the geographical vs. semantic location privacy plane. The darker the plot gets, the more data points there are in the corresponding geographical and semantic intersections. We exclude the scenario where the semantic tag of the events is always disclosed, because semantic location-privacy is always 0 in this scenario, hence it does not contribute to these plots. These figures present the change in the relationship between the geographical and semantic location-privacy. The obvious change occurs in the semantic dimension, though the change in the geographical location-privacy is non-negligible as well.

V. Related Work

A large amount of work has been devoted to quantifying location privacy, in particular when extra information (*i.e.,* different from location information *e.g.,* co-locations and location semantics) is available to the adversary. [2] is one of the first papers to identify and study inference attacks on location traces. Another notable example, on which our work is partially built, is presented in [13], [3]. In these papers, the authors propose a formal framework to quantify users' location-privacy when some (obfuscated) location information is available to the adversary. Their proposed framework relies on hidden Markov models for the location inference process and uses the expected error of the adversary as a metric for location privacy. The work presented in this paper enriches this framework by incorporating the rich semantic information increasingly disclosed by users on social networks. Similarly, but orthogonal, to our work, in [14], the authors study the effect of co-location information (*e.g.,* Alice and Bob are at the same (unknown) location at 2pm) on users' location privacy. As for obfuscation mechanisms, a detailed survey can be found in [1].

On the front of location semantics, several works study the



(a) Geo. Background (b) Geo. & Sem. Background

Fig. 12: Geographical location privacy vs. semantic location privacy. Note that we excluded the case of 'Sem. (actual, 0)' since it does not contribute to this plot (*i.e.,* it always results in no semantic location privacy).

semantic dimension of location information (some of them in the context of privacy). Several works, including [4], [5], [6] and [7], address the problem of identifying the points-of-interest (POIs) users visit, based on location traces. Some works extend existing location privacy metrics and definitions to take semantics into account. For instance, in [6], the authors propose a location-cloaking technique that ensures that the reported areas have a high semantic diversity in terms of the number of distinct venue types in the area. In [8], the authors propose the PROBE framework for implementing efficient, semantic-aware and personalized location cloaking. The concept of semantic diversity was originally formalized as l-diversity in [9] followed by related models including p-sensitivity [15], location diversity [16] and t-closeness [17]. Similarly, in [18], the authors extend the concept of *geo-distinguishability*, which applies differential privacy to location privacy [19], to take into account the semantic diversity of the reported locations. In [20], the authors propose the notion of C-safety, which not only takes into account semantics but also the sensitivity (in terms of privacy) of the different venue types. Using a taxonomy of venue types, the authors propose an efficient semantic-aware obfuscation mechanism. Finally, in [10], the authors study the implications of geographical and semantic obfuscation (through generalization) of users' check-ins on their perceived utility; in the evaluation of our work, we make use of the predictive model proposed in this paper.

Our work distinguishes itself from existing works as it incorporates semantic information in the *inference* process to better recover the users' locations, thus demonstrating the sensitive nature and the associated privacy risks of semantic information.

VI. Conclusion & Future Work

In this paper, we have investigated the effects of location semantics on geographical location-privacy of mobile users. We have considered two essential scenarios, specifically the case when an adversary, without knowing the semantic mobility patterns of the users, exploits the publicly available semantic information on locations, and secondly the case when the adversary knows the semantic mobility patterns of the users, in addition to knowing the location semantics. We have modeled the adversary that is aware of location semantics by using Bayesian networks and demonstrated that disclosing any level of semantic information on the visited locations improves his success. We have also studied and evaluated users' semantic location-privacy in the same context and shown that the semantic location privacy is diminished whenever the adversary has knowledge on users' semantic mobility patterns. Considering the increased amount of connectivity and huge data dissemination by individuals nowadays, this kind of knowledge is easy to obtain for any kind of digital adversary (especially as a service provider).

In summary, both the geographical and semantic location-privacy are at greater risk than revealed before, due to the multidimensional nature of location. When designing privacy-protection mechanisms, our aim must be to protect location privacy on a multidimensional scale, *i.e.,* considering the types of locations. Furthermore, the user mobility patterns also have an impact on both geographical and semantic location-privacy. Static protection mechanisms that do not take into account user history can fail to protect location privacy. For future work, we are planning to develop privacy-protection mechanisms that protect geographical and semantic location-privacy in a joint way and adapt their protection by using user history (an adaptive approach has been shown to protect geographical location-privacy better in a previous work by Agir *et al.*[21]). Furthermore, we believe that people have similar behavior patterns. For example, students regularly go to school in the morning and early afternoon, many people go to work in the morning and return home in the early evening, have lunch at around 12 o'clock, *etc.* Therefore, we would like to analyze the effect of the *collective* semantic mobility patterns on location privacy and reveal if just knowing the mobility patterns extracted from the community could help an adversary gain considerable improvements in his inference.

VII. Acknowledgments

The authors are thankful to Joana Machado for her help in building the utility predictive model. Parts of this work have been conducted while Kévin Huguenin and Reza Shokri were with EPFL, Lausanne, Switzerland. This work was partially funded by the Swiss National Science Foundation with grant 200021-138089 and the OpenSense2 project, funded by Nano-Tera.ch and financed by the Swiss Confederation.

References

[1] J. Krumm, "A survey of computational location privacy," *Personal Ubiquitous Comput.*, vol. 13, no. 6, pp. 391–399, Aug. 2009.

[2] ——, "Inference attacks on location tracks," in *Pervasive Computing*, vol. 4480, 2007, pp. 127–143.

[3] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Proc. of IEEE Symposium on Security and Privacy (S&P)*, 2011, pp. 247–262.

[4] J. Krumm and D. Rouhana, "Placer: Semantic place labels from diary data," in *UbiComp'13: Proc. of the 2013 ACM Int'l Joint Conf. on Pervasive and Ubiquitous Computing*, 2013, pp. 163–172.

[5] H. Liu, B. Luo, and D. Lee, "Location type classification using tweet content," in *ICMLA'12: Proc. of the 11th Int'l Conf. on Machine Learning and Applications*, vol. 1, 2012, pp. 232–237.

[6] B. Lee, J. Oh, H. Yu, and J. Kim, "Protecting location privacy using location semantics," in *KDD'11: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1289–1297. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020602

[7] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson, "The where in the tweet," in *CIKM'11: Proc of the 20th ACM Int'l Conf. on Information and Knowledge Management*, 2011, pp. 2473–2476.

[8] M. L. Damiani, E. Bertino, and C. Silvestri, "The PROBE Framework for the Personalized Cloaking of Private Locations," *Transactions on Data Privacy*, pp. 123–148, 2010.

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy Beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2007.

[10] I. Bilogrevic, K. Huguenin, S. Mihaila, R. Shokri, and J.-P. Hubaux, "Predicting Users' Motivations behind Location Check-Ins and Utility Implications of Privacy Protection Mechanisms," in *Proc. of the Network and Distributed System Security Symposium (NDSS)*, 2015, pp. 1–11.

[11] "Bayesian belief network package," accessed: 2015-08-16. [Online]. Available: https://github.com/eBay/bayesian-belief-networks

[12] F. V. Jensen, "Junction trees and decomposable hypergraphs." Judex Datasystemer, Aalborg, Denmark., Tech. Rep., 1988.

[13] R. Shokri, G. Theodorakopoulos, G. Danezis, J.-P. Hubaux, and J.-Y. Le Boudec, "Quantifying Location Privacy: The Case of Sporadic Location Exposure," in *The 11th Privacy Enhancing Technologies Symposium (PETS)*, 2011.

[14] A.-M. Olteanu, K. Huguenin, R. Shokri, and J.-P. Hubaux, "Quantifying the Effect of Co-locations on Location Privacy," in *PETS'14: Proc. of the 14th Int'l Symp. on Privacy Enhancing Technologies*. Amsterdam, The Netherlands: Springer, 2014, pp. 184–203.

[15] Z. Xiao, J. Xu, and X. Meng, "p-Sensitivity: A Semantic Privacy-Protection Model for Location-based Services," in *Proc. of International Conference on Mobile Data Management Workshops (MDMW)*, 2008.

[16] M. Xue, P. Kalnis, and H. K. Pung, "Location Diversity: Enhanced Privacy Protection in Location Based Services," in *Proceeding of International Symposium on Location and Context Awareness (LOCA)*, 2009.

[17] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE'07: Proc. of the IEEE 23rd Int'l Conf. on Data Engineering*, 2007, pp. 106–115.

[18] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "Constructing elastic distinguishability metrics for location privacy," in *PETS'15: Proc. of the 14th Int'l Symp. on Privacy Enhancing Technologies*, 2015.

[19] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *CCS'13: Proc. of the 2013 ACM SIGSAC Conf. on Computer and Communications Security*, 2013, pp. 901–914.

[20] A. Monreale, R. Trasarti, D. Pedreschi, C. Renso, and V. Bogorny, "C-safety: A framework for the anonymization of semantic trajectories," *Trans. Data Privacy*, vol. 4, no. 2, pp. 73–101, Aug. 2011.

[21] B. Agir, T. G. Papaioannou, R. Narendula, K. Aberer, and J.-P. Hubaux, "User-side Adaptive Protection of Location Privacy in Participatory Sensing," *Geoinformatica*, vol. 18, no. 1, pp. 165–191, 2014.