Berker Ağır*, Kévin Huguenin, Urs Hengartner, and Jean-Pierre Hubaux

# On the Privacy Implications of Location Semantics

**Abstract:** Mobile users increasingly make use of location-based online services enabled by localization systems. Not only do they share their locations to obtain contextual services in return (e.g., 'nearest restaurant'), but they also share, with their friends, information about the venues (e.g., the *type*, such as a restaurant or a cinema) they visit. This introduces an additional dimension to the threat to location privacy: location semantics, 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 the venue a user visits can be inferred, which also threatens her *semantic* location privacy. In this paper, we formalize this problem and analyze the effect of venue-type information on location privacy. We introduce inference models that consider location semantics and semantic privacy-protection mechanisms and evaluate them by using datasets of semantic check-ins from Foursquare, totaling more than a thousand users in six large cities. Our experimental results show that there is a significant risk for users' semantic location privacy and that semantic information improves inference of user locations.

**Keywords:** Location Privacy, Semantics, Inference, Social Networks

DOI 10.1515/popets-2016-0034
Received 2016-02-29; revised 2016-06-02; accepted 2016-06-02.

## 1 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 enjoy location-based services and location-based social networks. Users of such systems provide location information to the service providers in return for useful information, such as the location of the nearest restaurant, cinema or nearby friends, or simply to keep their friends posted about their activities. 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 benefiting from such services is her location data, which is detrimental to her privacy. This problem has been extensively investigated by the research community, focusing 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. Foursquare, for instance, lets its users check-in at specific nearby venues (selected from the Foursquare database of registered and confirmed venues, e.g., 'Super Duper Burger' in San Francisco), attach pictures and messages to their check-ins and report co-location with other users. Such location check-ins by themselves contain geographical information but also semantic information: For instance, the aforementioned venue is located at '2304 Market St' and is tagged as 'Burger Joint', which is a sub-category of 'Restaurant', which itself is a sub-category of 'Food' in Foursquare categories (see Figure 2). Hence, the approach to location privacy from a purely geographical perspective is not sufficient anymore. Additional dimensions of information about the activity of users can be exploited by service providers, thus reducing the effectiveness of existing privacy-protection mechanisms and threatening users' privacy. First, semantic information serves as additional location information: Knowing that a user is in a restaurant reveals some information about her location. Second, semantic information, combined with location information, can be exploited by learning patterns at the semantic level (e.g., people go to cinemas after going to restaurants). Such patterns are already available to (and used by) Foursquare, which makes next-venue recommendations to its users, e.g., "Places people like to go after 'Super Duper Burger': 'Castro Theatre (Movie Theatre, 429 Castro St)'" (see Figure 2).

***Corresponding Author: Berker Ağır:** EPFL, Lausanne, Switzerland, E-mail: berker.agir@epfl.ch
**Kévin Huguenin:** LAAS-CNRS, Université de Toulouse, CNRS, E-mail: kevin.huguenin@laas.fr
**Urs Hengartner:** University of Waterloo, ON, Canada E-mail: urs.hengartner@uwaterloo.ca
**Jean-Pierre Hubaux:** EPFL, Lausanne, Switzerland, E-mail: jean-pierre.hubaux@epfl.ch

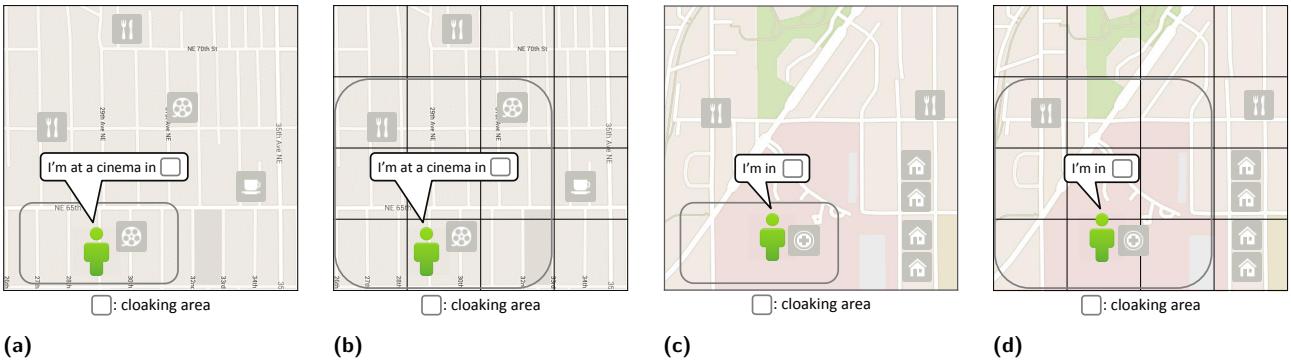



**Fig. 1.** Illustrative examples of the privacy threat caused by location semantics.(a) A user reports that she is in the depicted cloaking area and also that she is at a cinema. Her location can be easily pinpointed as there is only one cinema in the user's reported cloaking area, and this cinema occupies a small area compared to the cloaking area. The situation depicted in (b) demonstrates how the issue illustrated in (a) can be reduced by enlarging the cloaking area to include another cinema. An adversary can still narrow down the set of possible locations in the cloaking area, but now there are two locations with the tag cinema. (c) A user at a hospital reports a cloaking area without revealing her semantic information. As the hospital occupies a large proportion of the cloaking area, an adversary can infer that she is at a hospital, thus threatening the user's semantic location privacy. The situation depicted in (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 to address the issue illustrated in (c).

Figure 1 depicts two examples where the semantic dimension (i.e., the venue type) of a location can be exploited to infer the actual location and where 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. Because there is only one cinema in this cloaking area, one can easily pinpoint the user. In another example, depicted 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 with high confidence.

In this paper, we consider the case where users disclose not only their (obfuscated) geographical locations but also the (obfuscated) types of venue they are at in the form of check-ins on social networks, e.g., "Restaurant, downtown San Francisco". Previous studies have shown that such a level of information is often sufficient for the users [4], except, of course, when the purpose of the check-in is to become the "mayor" of the venue. Being able to report such obfuscated information would require making some modifications on the service. For instance, users could obfuscate their Foursquare check-ins and re-post an obfuscated version of them in a textual message on another social network (e.g., Twitter). Another solution would be that the service provider returns a list of venues to a user based on her coarse-grained location and lets her select a coarse-grained semantic information (e.g., "Food and beverage"). Last but not least, the techniques presented in this work can help build a tool for warning the users about the privacy risks associated with their check-ins. We focus on the semantic dimension of location check-ins and study its effects on location privacy, both at the geographical and semantic levels. Note that we consider a rather weak adversary (see Sections 2.3 and 3.1 for a detailed description of the adversarial model); therefore, the results presented in the paper constitute a lower bound in terms of the privacy loss stemming from semantic information.

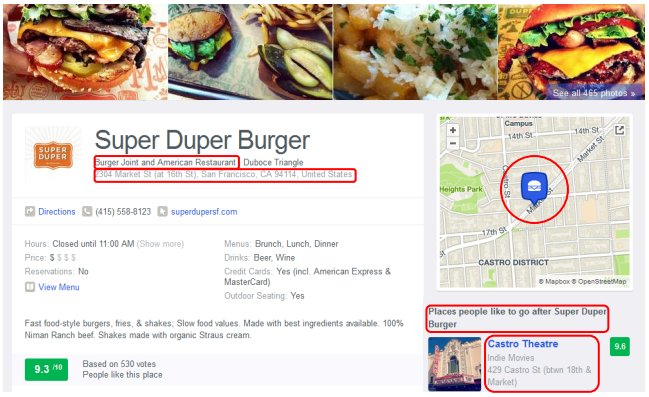



**Fig. 2.** Illustration of the information available to location-based social networks such as Foursquare: geographical (i.e., address) and semantic (i.e., venue category) information, semantic mobility profiles (i.e., 'Places people like to go after...'), *etc.* The most relevant pieces of information are circled in red.

To the best of our knowledge, our work is the first to confront, through data-driven experimentation, semantic information and semantic-aware location privacy protection mechanisms with a practical attack conducted by a concrete adversary. In a nutshell, we formalize the problem and build specific Bayesian networks to model users' behavior on which an adversary runs his inference attacks and we experimentally evaluate both geographical and semantic location privacy under such an adversarial model. In our experiments, we use the semantically-annotated location traces composed of Foursquare check-ins (collected through Twitter's public stream) of hundreds of users distributed across six large cities in North America and Europe. We also rely on a predictive utility model for obfuscated Foursquare check-ins [4]. We show that disclosing information about the type of visited locations, i.e., semantic location-information, decreases geographical location privacy by more than 50% (see Figure 5). For instance, in the extreme case where users disclose the precise type of venue they are at, their location privacy drops by 55% (from 420 m to 190 m). We also present the threat on semantic location privacy that deteriorates quickly as the adversary gains background information on user-mobility profiles, that are easy to build by crawling data publicly available 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 introduce the reader to the context and define the system model in Section 2. In Section 3, we present our adversarial model, inference approach and describe how we measure privacy. We explain our experimental setup and the datasets in Section 4 and report the evaluation results. In Section 5, we discuss the limitations of our approach and we propose improvement as future work. In Section 6, we survey related work. Finally, we conclude the paper and discuss future work in Section 7.

# 2 Background and System Model

We consider mobile users equipped with smartphones that have localization capabilities and Internet connectivity. These users move in a geographical area and make use of location-based online services. 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. In this setting, we consider an honest-but-curious service provider that is interested in inferring, based on his observations, users' actual geographical locations and the semantic tags associated with them, if any. Table 1 lists the notations used in the paper. Our model is built on top of Shokri et al.'s [3]; we detail the differences in Section 6.

## 2.1 Users

Mobile users with GPS-equipped connected devices move in a given geographical area that is partitioned into $M$ non-overlapping geographical regions/cells $\mathcal{R} = \{R_1, R_2, \dots R_M\}$. Geographical regions 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 regions in $\mathcal{R}$ contain venues annotated with semantic tags from the set $\{S_1, S_2, \dots, S_K\}$, i.e., a predefined list of categories (e.g., Foursquare defines such a list, organized as a tree,[1] and all registered venues are tagged with such a category). 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 $\{S_1, S_2, \dots, S_K\} \cup \{\perp\}$ to cover all semantic cases. Moreover, we consider discrete time instants over a limited-time period $\{1, \dots, T\}$. Note that the notion of venue types was introduced in the work of Shokri et al. [5].

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 $t \in \{1, \dots, T\}$, she generates an event consisting of her actual geographical region $r \in \mathcal{R}$ and a corresponding semantic tag $s \in \mathcal{S}$. This user event at time $t$ 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 $a_u = \{a_u(1) \dots a_u(T)\}$ the whole trace of user $u$.

**1** https://developer.foursquare.com/categorytree. Last visited: sep. 2015

**2** $\mathcal{P}$: Power set.

**Table 1.** Table of Notations.

| | |
|---|---|
| $\mathcal{R}$ | Set of geographical regions |
| $\mathcal{S}$ | Set of semantic tags |
| $a_u(t) = (r_t, s_t)$ | User $u$'s actual location at time instant $t$, where $r \in \mathcal{R}$ and $s \in \mathcal{S}$ |
| $o_u(t) = (r'_t, s'_t)$ | User $u$'s obfuscated location at time instant $t$, where $r' \in \mathcal{P}(\mathcal{R})^2$ |
| $a_u$ | Actual trace of user $u$ |
| $o_u$ | Obfuscated trace of user $u$ |
| $\mathbf{R}_t, \mathbf{R}'_t$ | The actual and obfuscated geographical location variables for time $t$ |
| $\mathbf{S}_t, \mathbf{S}'_t$ | The actual and obfuscated semantic location variables for time $t$ |
| $h_u(r, r', s, s')$ | A PPM modeled as a probability distribution function (PDF) employed by user $u$ (decomposed into $f_u(r, r')$ and $g_u(s, s')$) |
| $q_g, q_s$ | The PDF output by the inference attack |
| $\mathbf{dist}^G(\cdot,\cdot)$, $\mathbf{dist}^S(\cdot,\cdot)$ | Geographical and semantic distance metrics used for quantifying privacy |
| $\mathbf{GP}_u(t), \mathbf{SP}_u(t)$ | User $u$'s geographical and semantic location privacy at time $t$ |

## 2.2 Privacy Protection Mechanisms

For privacy reasons, users employ privacy-protection mechanisms (PPMs) before reporting their location and semantic information to an online service provider[3] Their privacy goal is to prevent the adversary from inferring at what geographical location and in what type of venue they are at. Typically, a PPM, 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 cloak), 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., report a parent tag of the venue's actual tag, w.r.t. a tag hierarchy, e.g., replace 'Burger joint' with 'Restaurant'[4] or 'Food') 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 $o_u = \{o_u(t) = (r', s')\}, \forall t \in \{1 \dots T\}$ of user $u$. We model a PPM as a probability distribution function that maps actual events to obfuscated ones (note that in the case of generalization, the PPM is deterministic). Specifically, we denote by functions $h(r, r', s, s')$ the probabilities to generate the obfuscated location/semantic tag $r', s'$(i.e., $\mathbf{Pr}(r', s'|r, s)$) that constitute the obfuscated event $o_u(t) = (r', s')$ given the actual event $a_u(t) = (r, s)$. Note that the location of a user at a given time instant is obfuscated independently from the other time instants.

Finally, we do not consider collaboration between users to protect their privacy (and prevent loss of privacy from each other). In addition, we assume that users' events are not anonymized.

## 2.3 Adversary

The adversary we consider in this paper is typically a service provider or an external observer who has access to obfuscated traces of users. He has two main purposes: (1) locate users at specific time instants, and (2) identify the types of the locations a user visits at specific time instants, 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, as explained in Section 3. Note that the inference process described below also applies to other adversaries such as users' friends and third party services on which users' check-ins are reposted (e.g., Twitter). However, the amount and the granularity of the information that is available to them can be more limited.

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 the users' mobility/behavior. He is assumed to have access to users' (partial) past events that he exploits to build a mobility profile for each user $u$, on both the geographical and semantic dimensions. Essentially, a user's mobility profile represents the user's transition probabilities over successive

**3** In the remainder of the paper, we refer to the *online service provider* as the *service provider* or the *adversary* for short.

**4** Note that this is strictly equivalent to reporting the sets of all tags that are sub-categories of tag 'Restaurant'

time instants, i.e., between geographical regions and between semantic tags. Formally, such a mobility profile (under a first-order Markovian assumption) is the set of the probability distribution functions $\mathbf{Pr}(r|\rho)$, $\mathbf{Pr}(s|\sigma)$ and $\mathbf{Pr}(r|s)$, where $\rho$ and $\sigma$ represent the user's previous location and semantic tag (as explained in Section 3).

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

# 3 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 [6, 7] under the assumption that people follow a bi-modal Markovian mobility process[5] (along the geographical and semantic dimensions) which we describe below. 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.

## 3.1 Inference and Background Knowledge

We assume that the adversary uses the following simple behavioral user model in the inference process[6]: 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 3):

1. The adversary knows the users' geographical transition profile, i.e., the *geographical background*, and assumes that the users move to new locations primarily based on their current locations. The type of place they visit (i.e., semantic tags) depend only on their current 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 and semantic transition profiles, 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., her next activity, characterized by the semantic tag of the venue she visits next) 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).

For the sake of simplicity for our experimentation, from this point on, we assume that geographical and semantic information are obfuscated independently from each other, using two functions $f_u$ and $g_u$ respectively (note that it is straightforward to include such joint PPMs in our formalism). Joint PPMs could be used to avoid the situations where a user reports a set of geographical locations and a semantic tag such that only some of the reported locations contain a venue with this tag.

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

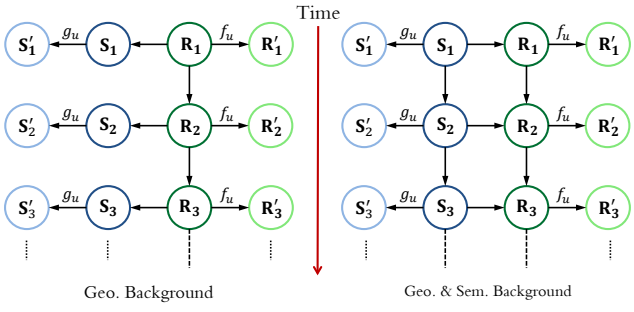



**Fig. 3.** The Bayesian networks representing the user models employed by the adversary. Nodes denote random variables and edges denote probabilistic dependencies between them (e.g., the arrow from $R_1$ to $R'_1$ corresponds to the obfuscation function $f_u$). The model on the left-hand side prioritizes geographical transitions with only geographical background known to the adversary. The model on the right-hand side prioritizes semantic transitions over geographical transitions with both geographical and semantic background. Protection mechanisms work separately on regions and semantic tags and they are independent.

**5** This means that a user's events at a given time instant only depend only on that user's event at the immediate past time instant.

**6** Note that the user traces we use in our experiments are real and are not generated from this model. Therefore, the fact that the considered user models rely on a set of simplifying assumptions limits the performance of the inference; as such, the experimental results presented in this paper constitute a lower bound of the privacy implications of semantic information.

### 3.1.1 Geographical-Only Background

As stated previously, the adversary has access to the users' geographical transition profile (built from past traces) in this scenario and carries out his attack by using (only[7]) this information as background information. He can correlate the sequential events of a user by using 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. 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 the semantic information disclosed by the users in his inference, even though he does not have any semantic background information.

This model is illustrated in Figure 3 (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 $\mathbf{S}'_t$ represent the 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 2.2. 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 powerful PPMs can be employed and used in this network, e.g., hiding the actual information completely with a given hiding probability.

The remaining conditional probabilities are those of 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. Essentially, $\mathbf{Pr}(\mathbf{S}|\mathbf{R})$ represents a uniform distribution over all semantic tags associated with a region $r$, e.g., if a region has 4 semantic tags 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 knowledge construction approach from [3].

### 3.1.2 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 to. Afterwards, she decides to go to a cinema, as she usually does after going to a restaurant. Considering her previous location, she picks the cinema that is most convenient for her. This model is depicted in Fig. 3 on the right-hand side.

The conditional probability distributions for the obfuscated events (i.e., $\mathbf{R}'_t$ and $\mathbf{S}'_t$) are the same as in the scenario with only the geographical background knowledge. The transitions between user events, however, 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 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. Therefore, 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 zero. This translates into the following formula:

**7** The purpose of considering such a limited adversary, used as a baseline, is solely to show the *inference power* of semantic background used in Section 3.1.2.

$$\mathbf{Pr}(\mathbf{R}_{t+1} = r|\mathbf{R}_t = \rho, \mathbf{S}_{t+1} = s) = \quad (1)$$

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

where $R_m$ denotes the set of regions that contain at least one venue with tag $s$ and $\alpha$ is a 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, $\alpha$ 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 $\alpha$ to 0.5, which we believe is a balanced treatment of user history.[8] Note that considering geographical and semantic background information separately 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.

## 3.2 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 algorithm is a probability distribution function (PDF) for each node in a given Bayesian network, i.e., the PDF $q_g$ over all regions at every time instant for user location and the PDF $q_s$ 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 $\mathrm{GP}_u(t)$ and $\mathrm{SP}_u(t)$, are computed as follows:

$$\mathrm{GP}_u(t) = \sum_{m=1}^{M} q_g(R_m, t) \cdot \mathrm{dist}^G(R_m, r), \quad (2)$$

$$\mathrm{SP}_u(t) = \sum_{k=1}^{K} q_s(S_k, t) \cdot \mathrm{dist}^S(S_k, s), \quad (3)$$

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

We use the Euclidean distance (in the projected coordinate system, i.e., Universal Transverse Mercator or UTM)[9] to compute the geographical distances between two regions by using the projected coordinates of their respective center points. We use the distance metric $d(\cdot)$ from graph-theory (i.e., the length of the shortest path between two nodes) on the category tree to compute the semantic distance between two tags, meaning that if two semantic tags are equal, then the distance is 0, if they have the same parent tag (e.g., 'American restaurant' and 'Burger joint' are both children categories of the 'Restaurant' category), the distance is 2, *etc.* We normalize the semantic distance between two tags by the sum of the tags' depths (i.e., the distance to the root).

$$\mathrm{dist}^S(s, s') = \frac{d(s, s')}{d(\text{'venue'}, s) + d(\text{'venue'}, s')}$$

This distance function takes into account the fact that, as one goes deeper in the tree, the graph-distance denotes a less significant semantic difference. For instance, "Italian Restaurants" and "American Restaurants" are not so different but "Food" and "Travel place" are.

# 4 Evaluation

We experimentally evaluate privacy on a real dataset of user traces composed of location check-ins that contain not only geographical location data but also semantic information in most cases (see Section 4.1). In

**8** We ran test experiments with different values of $\alpha$; we observed only small variations ($\sim 5\%$) of the median error, with better results for large values of $\alpha$ ($> 0.5$).

**9** Note that we did not take elevation into account in the computation of the geographical distance.

our experiments, we study the effects of location semantics on the geographical location privacy by comparing the privacy of users under a semantic-oblivious and a semantic-aware inference attack, in various configurations and with different PPM settings.

## 4.1 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 geographical and semantic information about the venues visited by the users of a large location-based social network. 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 the Amazon Mechanical Turk platform. This dataset was collected by the authors of [4] and made available online at https://homepages.laas.fr/khugueni/drupal/datasets. 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.

### 4.1.1 Location Traces with Semantics

Because we could not find large datasets of user check-ins with semantic information, we built our own dataset by running a data collection campaign through crawling. As a starting point, we use a tweet dataset we collected between January 2015 and July 2015 through Twitter's public stream. The dataset contains public geo-tagged tweets (i.e., Twitter lets users to attach their GPS coordinates to their tweets); we focused on six large cities: Boston (MA, USA), Chicago (IL, USA), Istanbul (Turkey), London (UK), New York (NY, USA) and San Francisco (CA, USA). We collected these tweets by identifying users through Twitter's public stream (i.e., ∼1% of the Twitter public timeline) and by fetching timelines of these users. A summary of the statistics of the dataset is provided in Table 2: We collected location check-in traces of a total of 1065 users. As we collected only public data and we neither interacted with the user nor inferred information not present in the dataset, IRB approval was not required.

The coordinates embedded in the geo-tagged tweets, however, do not contain semantic information (which we need for our evaluation). To obtain such information, we rely on Foursquare. Foursquare 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 an automatic text message 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[10] 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 (see Figure 13 of the Appendices, on page 182, for a snapshot of the tree), assigned to the venue. We used Foursquare's definition and implementation of location semantics *as is*. The results of our evaluation are dependent of the underlying semantic model; investigating alternative definitions of location semantics and other categorizations (e.g., from Facebook) is an interesting lead for future work. Because it uses semantic tags (organized as a tree) and because its main feature is to let users check-in at venues, Foursquare constitutes a perfect data source for our evaluation. Note that, unlike in works such as Krumm's [8] in which semantic

**Table 2.** Filtered Dataset Statistics

| City | Users | Tweets | Check-ins |
|---|---|---|---|
| Boston | 79 | 6,687 | 5,276 |
| Chicago | 136 | 14,248 | 11,755 |
| Istanbul | 196 | 22,203 | 17,005 |
| London | 239 | 18,685 | 15,018 |
| New York | 242 | 21,249 | 14,240 |
| San Francisco | 173 | 16,739 | 13,650 |

**Table 3.** Experimental Setup

| | |
|---|---|
| Number of iterations | 10 |
| Size of each area | $2.4 \times 1.6$ km ($12 \times 8$ cells) |
| Average Proportion of Foursquare tweets per user (i.e., tweets w/ semantic information) | 77% |

**10** Note that GPS coordinates in the tweets might slightly differ from registered venue coordinates at Foursquare. In such cases, we use the coordinates of the venues from Foursquare.

information is *inferred* from the users' location traces, we use only ground-truth semantic data extracted from the users' check-ins. We show the venue density and the Foursquare tweet density in the considered cities in Figure 14 (Appendices, p. 183), which shows a Foursquare venue heat map and a Foursquare check-in heat map.

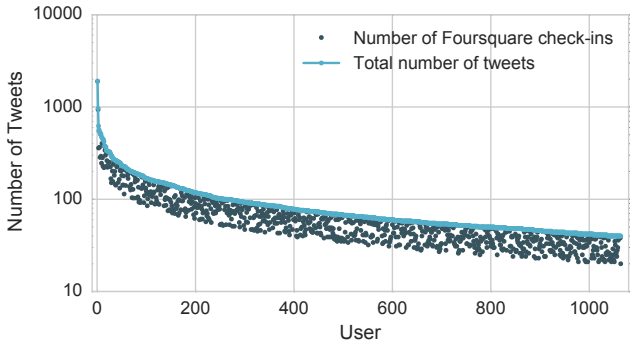



**Fig. 4.** Number of Foursquare check-ins/tweets and the total number of tweets per user (in decreasing order) in the filtered dataset used in our experiments (log-scale on the y-axis).

In our evaluation, we focus (due to computational limitations) on the tweets and check-ins in small geographical areas of size approximately 2.4×1.6 km around the cities of Boston, Chicago, Istanbul, London, New York and San Francisco. We define one such area around each of the six cities, and we divide each of them into 96 cells by using a regular grid of $12 \times 8$ cells (each of size $200 \times 200$ m). We determine the most dense such areas and extract users with at least 40 tweets in each region. We further filter out users whose Foursquare tweets (i.e., check-ins) account for 1]less than 50% of all their tweets (i.e., most of the tweets used in the experiments contain venue information). The final dataset contains a total of 1065 users (57% male, 41% female, 2% unknown); see Table 2 for detailed statistics and Figure 4 for users' count of Foursquare and total tweets. We included all the tweets of a user in the knowledge construction of the adversary and for each user we use a randomly selected sub-trace of length 5 in each experiment. There are 10,970 venues in our filtered dataset and the tag distribution over these venues is shown in Figure 12 (Appendices, p. 182).

**Dissemination of the dataset**

Although the terms and conditions of Twitter[11] and Foursquare[12] prevent us from making the dataset directly available for automatic download as we need to make sure that the requesting party agrees to comply with the aforementioned terms, we will be happy to provide our dataset (and the script used for collecting data) to other researchers upon request.

The dataset contains all the considered check-ins, each of which is characterized by a timestamp, a user id, a geographical location (as reported in the tweet), a geographical location (as reported in the Foursquare venue information), and the Foursquare venue type in the form of a tag, and will be made available in the csv file format. It will also contain a snapshot of Foursquare category tree at the time of data collection.

### 4.1.2 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 dataset collected and made available by the authors of [4]. The fact that the survey focuses on Foursquare check-ins makes it perfectly adequate for our dataset and hence for our evaluation. 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, 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 5 (Appendices, p. 183, extracted from the original

**11** https://dev.twitter.com/overview/terms/agreement-and-policy

**12** https://developer.foursquare.com/overview/venues

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 [4] (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).

## 4.2 Experimental Setup

**Methodology:**
We partitioned each of the six considered areas (one for each city considered in the dataset) into 96 cells, each identified by an ID, using an $12 \times 8$ regular square grid. We then mapped the locations in the users' traces to the corresponding region IDs, and we kept the semantic tag. We implemented our Bayesian network-based models in Python by using the Bayesian Belief Networks library provided by eBay [9]. We applied certain protection approaches (listed below) on the users' traces, obtaining *protected*/observed traces that our Bayesian networks use as observations, and applied the junction-tree inference algorithm [10] which achieves optimal inference. The output of the inference algorithm is a probability distribution function for each unknown (inferred) variable, which we use in our privacy metrics (see Equations (2) and (3)).

**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 3.1):

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 Section 3.1. The transition probabilities are estimated from the number of geographical transitions in the whole traces of users.
2. **Geographical and Semantic Background:** The adversary is assumed to have more knowledge 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.

In many cases, such information can be obtained by the service provider. In cases where only little background information about individual users is available, the service provider can aggregate data across users with similar profiles.

**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. As mentioned above, joint protection mechanisms could be used for improved performance; we leave the design of such mechanisms to future work.

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 the hiding probability $\lambda$). 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 $2 \times 2$ obfuscation: ($i$) with probability $1-\lambda$, generates an obfuscation area consisting of 4 adjacent regions/cells, one being the actual location of the user, or ($ii$) with probability $\lambda$, 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. From a privacy perspective, this constitutes a worst-case scenario.
2. **Parent-tag obfuscation**. This is a generalization based on the semantic tag tree derived from Foursquare's category hierarchy. In this case, 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. It has been shown, for Foursquare check-ins, that reporting the parent tag of a venue is often sufficient to meet the purpose of the original check-in [4].
3. **Parent-tag obfuscation with hiding**. 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 $\lambda$.
4. **Complete hiding of semantic tags [baseline]**. In this case, we never disclose semantic tags. This corresponds to a pure geographical approach (as taken in previous works); as such it constitutes our baseline.

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

## 4.3 Experimental Results

In this section, we analyze the experimental results with different protection mechanisms in various settings.

### 4.3.1 Effect of Semantic Information on Location Privacy

We first analyze 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 $\lambda$. Specifically, given a geographical obfuscation parameter (e.g., $2\times2$ obfuscation), for each $\lambda$ we evaluate four different semantic protection approaches (explained in Section 4.2) that are employed together with the obfuscation mechanism.

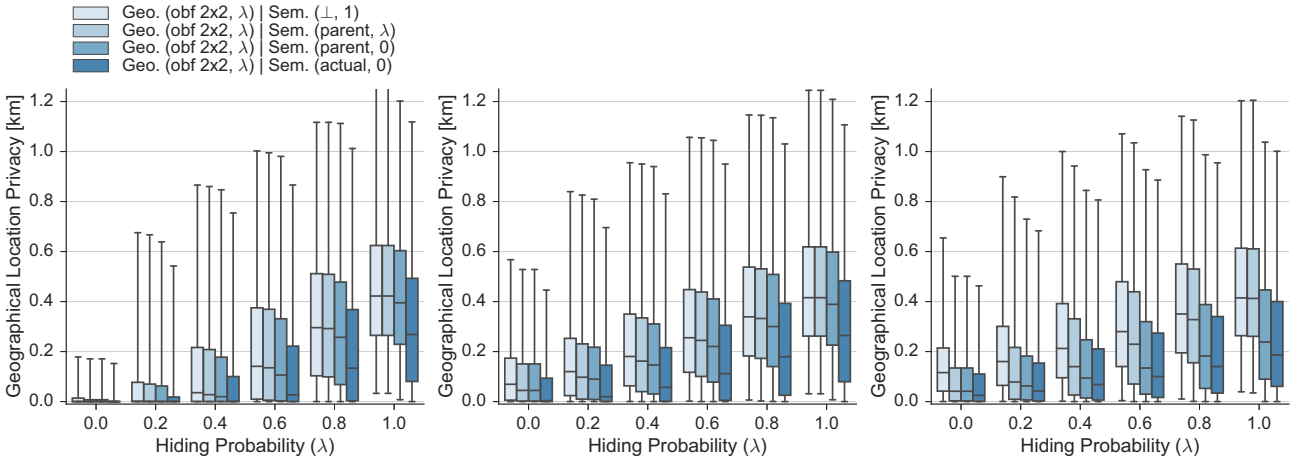
We present the results in Figure 5, where the x-axis represents the hiding probability $\lambda$ (used for geographical obfuscation and parent-tag semantic generalization) and the y-axis represents the geographical location privacy in kilometers (i.e., the distance between a user's actual discretized location and that inferred by the adversary, as described in Equation (2)). A privacy of a few hundreds of meters (typically a city-block) provides a reasonable protection against precise localization/tracking and limits the possibility to infer the exact place a user visits or her exact address. We plot the geographical location privacy aggregated over all users, all events and all iterations of simulations for each protection mechanism and hiding probability ($\lambda$) pair using box plots. These box plots show the 1st, 2nd, 3rd quartiles of the data and the 98% confidence intervals.

We consider four scenarios (geographical obfuscation and semantic generalization) and plot the corresponding results, e.g., "Geo. (obf $2\times2$, $\lambda$) | Sem. (parent, $\lambda$)" means that (1) geographical locations are hidden with probability $\lambda$ and obfuscated by reporting $2\times2$ cloaking areas otherwise, and (2) semantic tags are hidden with probability $\lambda$ and generalized by reporting the parent tag otherwise; the darker a box-plot is, the higher the amount of disclosed information is. In our experiments, we employed both $2\times2$ and $4\times4$ cloaking.[13]

We observe that as we disclose more semantic information, along with the obfuscated geographical location (from left to right for each $\lambda$ value), the median location privacy consistently decreases in all cases. For instance, it can be observed in Figure 5c ($\lambda=1$), that disclosing the actual semantic tag decreases the median location privacy by 55% (from 420 m to 190 m) and disclosing the parent tag decreases it by 43%. Also, unsurprisingly, the privacy level increases as we increase the granularity of the location (i.e., from $2\times2$ obfuscation in Figure 5a to $4\times4$ obfuscation in Figure 5b). Note that for $\lambda=1.0$, the parent-tag generalization with hiding probability $\lambda$ 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 5.

We also analyze the effect of employing semantic background information (i.e., the histories of users' transitions between semantic tags) in the inference process, in addition to the geographical background information that is already employed in all our experiments. We compare the two scenarios where $4\times4$ geographical obfuscation with hiding probability $\lambda$ is used (i.e., Fig-

**13** We acknowledge the fact that $4\times4$ cloaking is rather large compared to the size of the grid. We intend to consider larger grids as part of future work.




**(a)** $2\times2$ Obfuscation w/o Sem. Background **(b)** $4\times4$ Obfuscation w/o Sem. Background **(c)** $4 \times 4$ Obfuscation w/ Sem. Background

**Fig. 5.** Geographical location privacy levels over different protection and learning scenarios.

ures 5b and 5c, with and without semantic background information respectively). We observe that, for instance in the 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 can also be observed 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 median accuracy by 10 to 115 meters when the users disclose some semantic information in their traces. This is clear in Figure 7, that shows the difference between Figures 5b and 5c (i.e., the information gain of the adversary between the two scenarios). The reason why the adversary gains more information in the case of parent-tag obfuscation compared to no semantic protection is that when users disclose their semantic tags, their privacy level is already lower; hence the potential information gain of the adversary in *with* semantic background scenario is naturally lower.

Figure 6 depicts the average geographical location privacy in each of the six considered cities (with and without semantic background, aggregated over all values of $\lambda$ and over the two sizes of the cloaking area). It can be observed that it is quite comparable among cities: Despite the difference in terms of culture and urban planning, we did not observe major differences across cities in terms of user privacy in the presence of semantic information. It can also be observed that semantic background information improves the performance of the inference, thus decreasing users' geographical location privacy. Note that our experiments include some randomness and as a result, in some situations (e.g., New York) the background information slightly misleads the adversary.

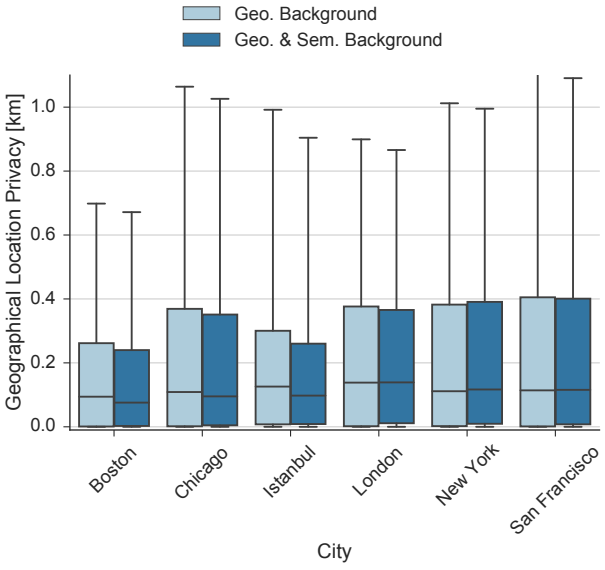



**Fig. 6.** Average geographical location privacy over all users in each considered city.

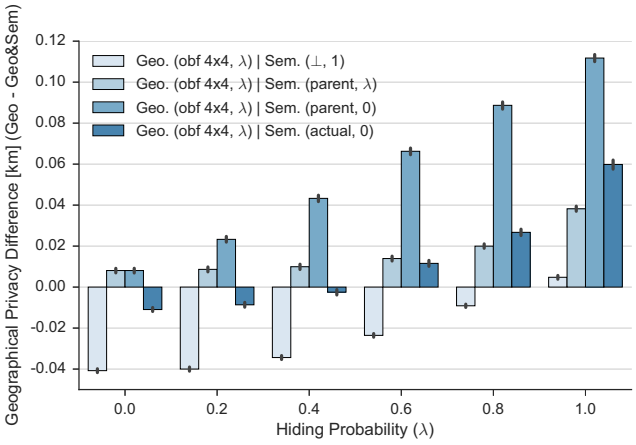



**Fig. 7.** Difference of geographical location-privacy levels between the cases *with* semantic background and *without* semantic background with $4 \times 4$ geographical obfuscation and varying $\lambda$. As soon as users disclose some semantic information, the performance of the inference increase when using semantic background information about users. Interestingly, when users completely hide the semantic tags of their locations, the adversary is less successful when he uses the semantic background information.

### 4.3.2 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. To comply with the experimental setup of [4], 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 4 and illustrated in Figure 8. We set the hiding probability $\lambda$ to 0.2.

**Table 4.** Description of the different obfuscation levels.

| Obfuscation | Description |
|---|---|
| **Ls-Lg** | **Semantic tag, $2 \times 2$ geographical region** |
| **Hs-Lg** | **Parent semantic tag, $2 \times 2$ geographical region** |
| **Ls-Hg** | **Semantic tag, $4 \times 4$ geographical region** |
| **Hs-Hg** | **Parent semantic tag, $4 \times 4$ geographical region** |

We plot the results in Figure 9. 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 [4]. This means that, as far as geographical location privacy is concerned, users should always prefer Ls-Hg over Hs-Lg. As for semantic location privacy (which we analyze in detail in the next sub-section), it can be observed that geographical obfuscation is quite beneficial as the use of high geographical obfuscation substantially increases the users' semantic location privacy at a cost of a small decrease in utility. In the case where low semantic obfuscation is used, the semantic location privacy is zero as the users reveal the actual semantic tags of their locations.

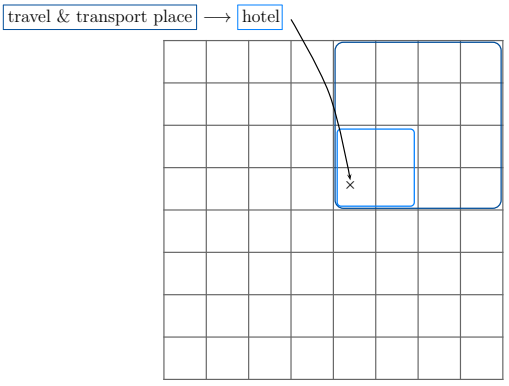



**Fig. 8.** Illustration of the obfuscation levels used in the experiments. Light blue frames denote low levels of obfuscations whereas dark blue frames denote high levels of obfuscation.

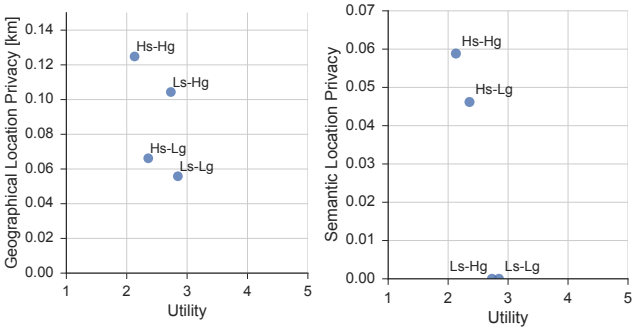
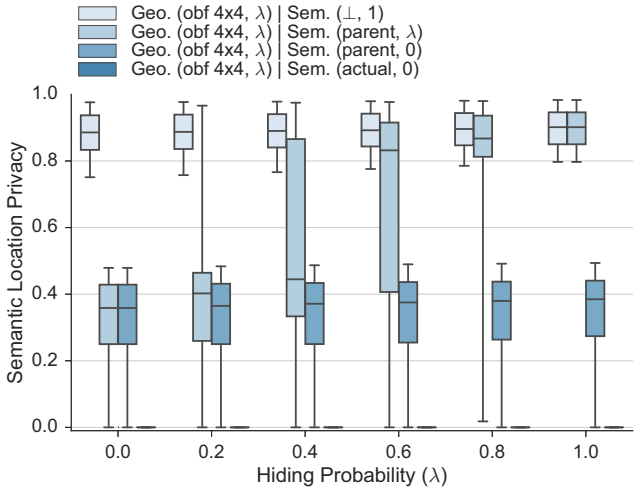


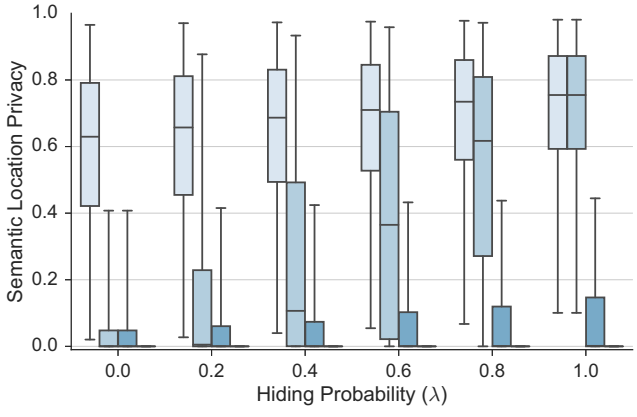

**Fig. 9.** Privacy vs. Utility in four different scenarios (Ls-Lg, Hs-Lg, Ls-Hg, Hs-Hg, for $\lambda = 0.2$).

### 4.3.3 Semantic Location Privacy

Finally, we evaluate the semantic location privacy and present the loss of privacy in the semantic dimension of location. As in the figures depicting geographical location privacy, 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.




**(a)** $4 \times 4$ Obfuscation w/o Sem. Background




**(b)** $4 \times 4$ Obfuscation w/ Sem. Background

**Fig. 10.** Semantic location privacy levels over different protection scenarios with geographical and semantic background knowledge of the adversary.

In Figure 10, we present the semantic location-privacy results for $4 \times 4$ obfuscation with hiding probability $\lambda$ in both 'Geographical background' and 'Geographical & Semantic Background' scenarios. In both cases (shown separately in figures 10a and 10b), as we protect the semantic information in 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 $\lambda$, 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 substantial as can be seen in Figures 10a and 10b: 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 80% in other protection scenarios.

## 5 Discussion

In this paper, we presented a semantic-aware location inference scheme, which we tested against several simple privacy-protection mechanisms (PPM), to prove that the threat on location privacy is more acute when the semantic dimension of location is taken into account. However, this is just a first step towards developing smarter PPMs, which take into account the semantic dimension of location privacy (together with the geographical dimension). The results we demonstrated in this work serve an important purpose: Understanding how to develop joint PPMs that protect geographical and semantic location privacy together and by taking into account user history and profiles. Our work enables evaluation of such PPMs by paving the way for testing and adapting them w.r.t. the success of the adversary in an adaptive manner as well as optimizing jointly privacy and utility. As part of future work, we plan to use this framework to develop smarter PPMs. For instance, we intend to consider PPMs such as "If the cloaking area contains only one Burger joint opened at the considered time instant, either increase the size of the cloaking area or use the parent semantic tag, depending on which option brings the lowest utility loss". Another option is to implement warning mechanisms that interferes with user actions whenever a user wants to check-in at specific sensitive places and warns the user of the privacy risks. Such an approach would increase the awareness of users in addition to providing sufficiently protected privacy levels.

A first limitation of this work is the fact that the adversary we considered uses a basic user behavioral model. As such, the results we present constitute a lower bound on the privacy loss: The adversary can actually strengthen his attack by increasing the complexity of the model he uses. For instance, he could exploit the temporal properties of locations and semantics: Users tend to have periodic routines (e.g., daily/weekly), such as staying home at night, going to work or school during the day and having lunch around noon, and venues have characteristic opening hours. By taking into account the time dimension, we could show that the threat is actually greater than what we demonstrate. Furthermore, the information we considered is in fact a subset of what a typical adversary (i.e., a service provider) can collect. The fact that the adversary has access to geographic and semantic profiles (i.e., background information) may be considered as rather strong. However, such knowledge can be built not only from obfuscated traces, but also by aggregating the data of several similar users, thus building more generic models (as done by Foursquare for next place recommendations).

A second limitation of this work is the size and the nature of the dataset: We considered "only" 1065 users (whom we have only little demographic information about) in six cities, who linked their Foursquare and Twitter accounts and made the tweets generated by Foursquare *public*. Such a sampling method could introduce a bias in the experimental results. Moreover, the grid size is rather small, especially as the obfuscated regions can be as large as 4×4. Finally, the time granularity of our dataset is somewhat coarse. As part of future work, we will work on increasing the size and quality of our user dataset and better characterizing the users it contains in order to make our experimental results more generalizable. It would also be interesting to use traces from location-based social networks (and the associated tag hierarchy) other than Foursquare.

# 6 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 [3, 11]. 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. Note that Shokri's framework can be used *as is* to include semantic information by defining a location as a couple (geographical location, semantic location). This however, makes existing techniques for background construction inefficient due to the sparsity of the transition data (although many transitions go from one geographical region to another, the number of transitions from a couple (region, semantic tag) to another is significantly reduced). Also, recent work have shown that moving from Hidden Markov Models to Bayesian networks enables the adversary to take into account more complex information such as co-location [12]. The main differences between our work and Shokri et al.'s are (1) the use of general Bayesian networks to model users' behavior and (2) a two-step background construction (i.e., first semantic, then geographical) to deal with sparse data. Similarly, but orthogonal, to our work, in [12], 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 semantic dimension of location information (some of them in the context of privacy). Several works, including [8], [13], [14] and [15], address the problem of identifying the points-of-interest (POIs) users visit, based on location traces. Unlike our work, these works do not consider semantic information *reported* by the users. Hence, obfuscating semantic information is not directly possible. Barak et al. propose an anonymization technique based on semantic cloaking, that is, replacing actual coordinates by *personal* semantic labels such as 'home' (by opposition to the *universal* labels we considered, such as 'restaurant') [16]. Some works extend existing location privacy metrics and definitions to take semantics into account. For instance, in [14], the authors propose a location-cloaking technique that ensures that the reported regions have a high semantic diversity in terms of the number of distinct venue types in the area. In [17], 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 [18] followed by related models including $p$-sensitivity [19], location diversity [20] and $t$-closeness [21]. Again, these works focus mostly on providing formal *semantic* location privacy guarantees by obfuscating *location information*, whereas our work considers both geographical and semantic information and investigates the privacy implications on both dimensions, based on statistical inference. Similarly, in [22], the authors extend the concept of *geo-distinguishability*, which applies differential privacy to location privacy [23], to take into account the semantic diversity of the reported locations. Differential privacy-based frameworks and inference-based frameworks are fundamentally different in their approach to privacy quantification. In [24], 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. 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.

Finally, complementary to our approach, in [4], 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.

In the general context of location sharing, a number of cryptographic protocols have been proposed (e.g., [25] for private and cheat-proof "mayorship"-badges, one of the main feature of location-based social networks, and [26, 27] for sharing location with friends without the service provider learning the users' locations). Such solutions, however, involve cryptographic operations and require technical modifications of the service. Related cryptographic protocols, which provide privacy-preserving features, are proposed in [28] and [29]. They rely on secure multy-party computations (garbled circuits) and homomorphic encryption schemes, respectively. Such approaches can be applied to, for instance, friend-finding applications without revealing user locations, but they require careful analysis and extension to incorporate the semantic dimension of location. These mechanisms aim to provide privacy-preserving features in specific applications, and it is not straightforward to modify them in order to cover the cases where people want to disclose their current activities, i.e., their location semantics.

# 7 Conclusion

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.

In summary, both the geographical and semantic location privacy are at greater risk than revealed before, due to the multidimensional nature of location data. 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, because we believe that people have similar behavior patterns, we intend to analyze the effect of *collective* mobility patterns on location privacy.

# 8 Acknowledgments

The authors are thankful to Joana Machado for her help in building the utility predictive model, to Alexandra-Mihaela Olteanu for her help with Bayesian networks, and to Reza Shokri for his feedback during the first stages of the project. Parts of this work have been conducted while Kévin Huguenin was with EPFL, Lausanne, Switzerland. This work was partially funded by the Swiss National Science Foundation with grant 200021-138089 and by the Swiss Confederation through the Nano-Tera OpenSense2 project.

# 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 the IEEE Symp. on Security and Privacy (S&P)*, 2011, pp. 247–

262.

[4] 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.

[5] R. Shokri, J. Freudiger, and J.-P. Hubaux, "A unified framework for location privacy," in *Proc. of the Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETS)*, 2010.

[6] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[7] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

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

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

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

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

[12] A.-M. Olteanu, K. Huguenin, R. Shokri, M. Humbert, and J.-P. Hubaux, "Quantifying interdependent privacy risks with location data," *IEEE Transactions on Mobile Computing*, p. 14, 2016, to appear.

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

[14] B. Lee, J. Oh, H. Yu, and J. Kim, "Protecting location privacy using location semantics," in *Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, 2011, pp. 1289–1297.

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

[16] O. Barak, G. Cohen, and E. Toch, "Anonymizing mobility data using semantic cloaking," *Pervasive and Mobile Computing*, 2015, to appear.

[17] 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.

[18] 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.

[19] 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.

[20] M. Xue, P. Kalnis, and H. K. Pung, "Location Diversity: Enhanced Privacy Protection in Location Based Services," in *Proc. of the Int'l Symp. on Location and Context Awareness (LOCA)*, 2009.

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

[22] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "Constructing elastic distinguishability metrics for location privacy," in *Proc. of the Privacy Enhancing Technologies Symp. (PETS)*, 2015.

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

[24] 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.

[25] B. Carbunar, R. Sion, R. Potharaju, and M. Ehsan, "The shy mayor: Private badges in geosocial networks," in *Proc. of the 10th Int'l Conf. on Applied Cryptography and Network Security (ACNS)*, 2012, pp. 436–454.

[26] C. Dong and N. Dulay, "Longitude: A privacy-preserving location sharing protocol for mobile applications," in *Proc. of the Int'l Conf. on Trust Management (IFIPTM)*, 2011, pp. 133–148.

[27] M. Herrmann, A. Rial, C. Diaz, and B. Preneel, "Practical privacy-preserving location-sharing based services with aggregate statistics," in *Proc. of the ACM Conference on Security and Privacy in Wireless (WiSec)*, 2014, pp. 87–98.

[28] B. Mood, D. Gupta, K. Butler, and J. Feigenbaum, "Reuse it or lose it: More efficient secure computation through reuse of encrypted values," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 582–596.

[29] G. Zhong, I. Goldberg, and U. Hengartner, "Louis, lester and pierre: Three protocols for location privacy," in *Privacy Enhancing Technologies*. Springer, 2007, pp. 62–76.

# Appendices

## Analysis of the effect of $\alpha$ (Eq. (1))

In this appendix, we present the preliminary results of our analysis of the effect of parameter $\alpha$ (used in the transition probabilities of the Bayesian model, see Eq. (1)), in the case where the adversary has access to both geographical and semantic background information. In Figure 11, we plot the geographical location privacy obtained for different values of $\alpha$ (with mixed hiding probabilities). We observe that with increasing $\alpha$ (i.e., prioritizing geographical information over semantic information), users obtain higher location privacy (i.e., the adversary is less successful) when they disclose the semantic tag. However, in the cases of hiding the semantic tags and parent-tag cloaking with hiding, the $\alpha$ value has less effect. This shows that an actual adversary could and should tune the model used in the attack, based on his observations, in order to improve the performance of the inference.

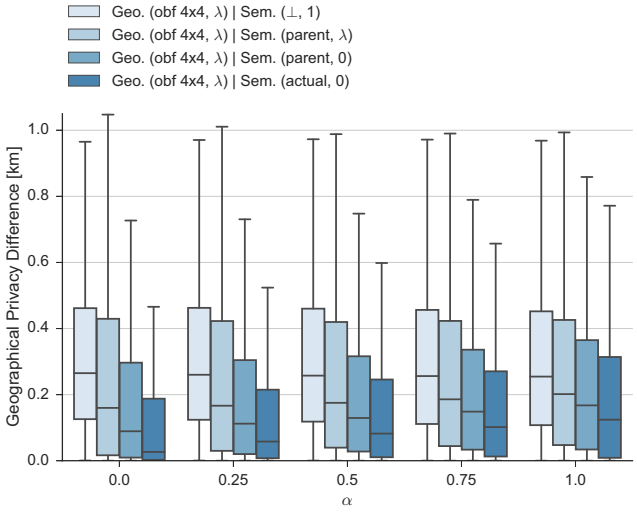



**Fig. 11.** Effect of parameter $\alpha$ on the users' geographical location privacy.

## Supplementary Figures

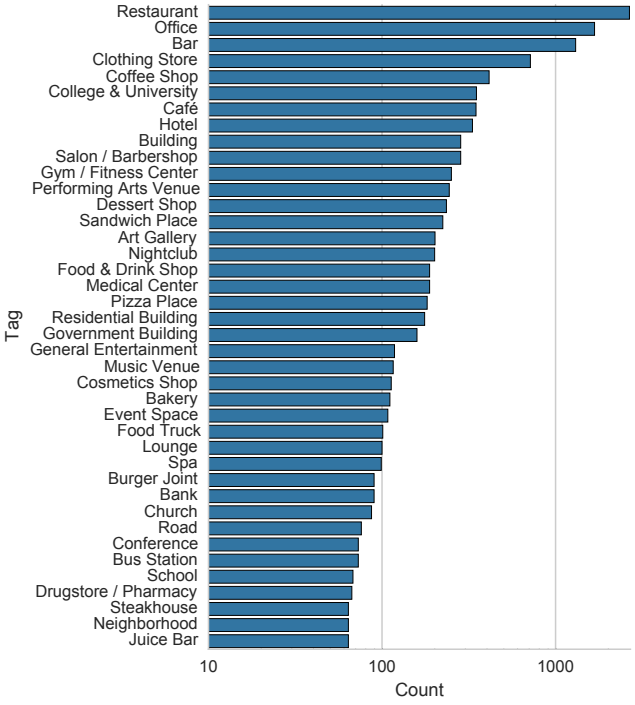



**Fig. 12.** Number of venues per semantic tag in the filtered dataset for the top 40 common tags (log-scale on the x-axis).

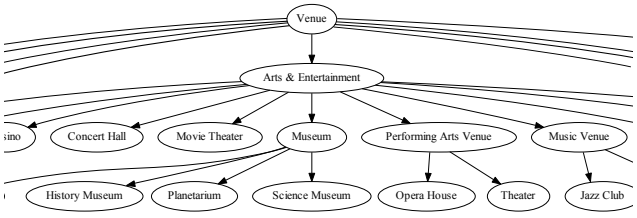



**Fig. 13.** Partial view of the Foursquare category hierarchy that we use as our semantic tag tree in our evaluation. The 'Venue' tag is the root of the category tree.

**Table 5.** Example of obfuscated check-ins with different combinations of geographical and semantic obfuscation (source: [4]).

| 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) |

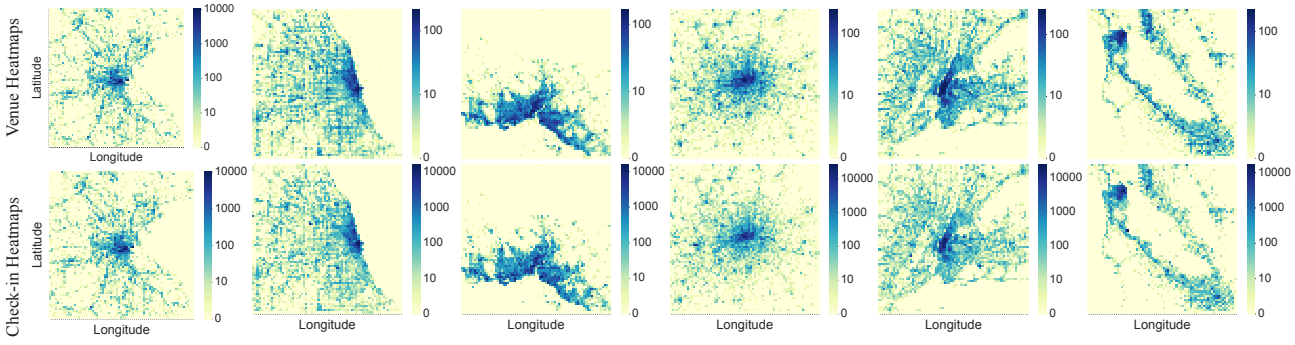



**Fig. 14.** Foursquare venue and check-in heat maps (i.e., count distribution) in six cities from the raw dataset.