A Predictive Model for User Motivation and Utility Implications of Privacy-Protection Mechanisms in Location Check-Ins

Kévin Huguenin, *Member, IEEE,* Igor Bilogrevic, Joana Soares Machado, *Member, IEEE,* Stefan Mihaila, Reza Shokri, Italo Dacosta, *Member, IEEE,* Jean-Pierre Hubaux, *Fellow, IEEE*

Abstract—Location check-ins contain both geographical and semantic information about the visited venues. Semantic information is usually represented by means of tags (e.g., "restaurant"). Such data can reveal some personal information about users beyond what they actually expect to disclose, hence their privacy is threatened. To mitigate such threats, several privacy protection techniques based on location generalization have been proposed. Although the privacy implications of such techniques have been extensively studied, the utility implications are mostly unknown. In this paper, we propose a predictive model for quantifying the effect of a privacy-preserving technique (i.e., generalization) on the perceived utility of check-ins. We first study the users' motivations behind their location check-ins, based on a study targeted at Foursquare users ($N = 77$). We propose a machine-learning method for determining the motivation behind each check-in, and we design a motivation-based predictive model for the utility implications of generalization. Based on the survey data, our results show that the model accurately predicts the fine-grained motivation behind a check-in in 43% of the cases and in 63% of the cases for the coarse-grained motivation. It also predicts, with a mean error of 0.52 (on a scale from 1 to 5), the loss of utility caused by semantic and geographical generalization. This model makes it possible to design of utility-aware, privacy-enhancing mechanisms in location-based online social networks. It also enables service providers to implement location-sharing mechanisms that preserve both the utility and privacy for their users.

Index Terms—Human factors; Location-based social networks; Utility; Privacy, Location semantics;

1 Introduction

Users of popular online social networks (OSNs), such as Facebook, Foursquare, and Twitter, are offered the possibility to share their location information with other users. Such a feature, commonly known as *location check-in*, enables users to report to their friends that they are in a particular venue (*e.g.*, a restaurant) and to provide recommendations and/or comments about it. Many users take advantage of this feature; it is estimated that around 30% of users attach locations to their posts [2]. The reason for sharing locations includes the desire to connect with users' social circles and to project an interesting image of themselves [3], [4], thus achieving an objective greater than simply disclosing geographical information [5], [6].

- *Manuscript received: August 16, 2017;*
- *This article is a revised and extended version of a paper that appears in the Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015) Bilogrevic* et al. *[1].*
- *Kévin Huguenin is with HEC Lausanne, UNIL, Lausanne, Switzerland (e-mail: kevin.huguenin@unil.ch). This work was partially carried out while the author was with EPFL, Lausanne, Switzerland.*
- *Igor Bilogrevic is with Google, Zurich, Switzerland (e-mail: ibilogrevic@google.com). This work was carried out while the author was with EPFL, Lausanne, Switzerland.*
- *Reza Shokri is with the computer science department at NUS, Singapore (e-mail: reza@comp.nus.edu.sg). This work was carried out while the author was with ETH, Zurich, Switzerland.*
- *Stefan Mihaila, Joana Soares Machado, Italo Dacosta, and Jean-Pierre Hubaux are with EPFL, Lausanne, Switzerland (e-mail: stefan.mihaila@epfl.ch; joana.machado@ieee.org; italo.dacosta@epfl.ch; jean-pierre.hubaux@epfl.ch).*

By checking-in on so-called location-based social networks (LBSNs) about a place or an event, such as a restaurant or a gathering, users implicitly accept to reveal the geographical coordinates and the semantic information of the place. For example, when they check in at a restaurant, users reveal the exact location of that restaurant, as well as its type or category, represented in the form of tags, such as "burger joint", as illustrated in Figure 1 (venue types are usually selected from a pre-defined set of tags, organized as a hierarchical tree, where the "burger joint" tag could be a descendant of the "restaurant" tag.). This can lead to the exposure of additional private information beyond what they intended to share and can make inference attacks more powerful [7], much to the detriment of the users' privacy. Typical adversaries include other individuals (users of the social network, social contacts of the user who shared the information) and service providers (the social network operator or third-parties); potential threats include discrimination of all sorts. For instance, a location-based social network provider can exploit semantic information to learn activity patterns (e.g., people go to bars after going to restaurants)[1] to better infer the locations of its users and to profile them. A collection of location check-ins by a set of users can lead to their re-identification and also to the inference of more personal information (e.g., complete location trace, co-travelers, activities) [8], [9], [10]. The risks are even higher when users share semantic information

1. Foursquare analyzes such patterns across all its users and uses them to make next-venue recommendations, as illustrated in Figure 1.

as well. Several works have studied the privacy risks and the associated protection mechanisms for location semantic data [11], [12], [13].

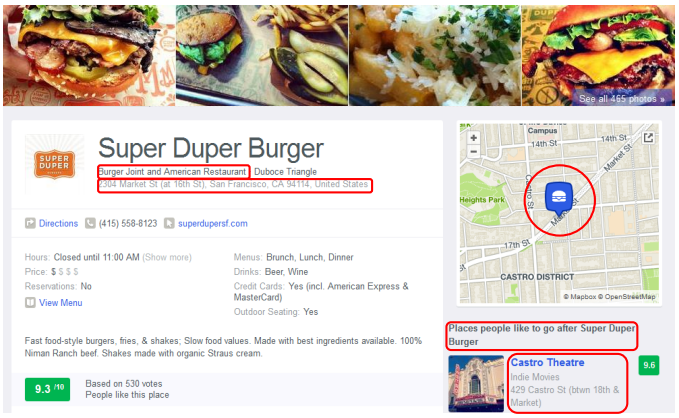

Figure 1. Illustration of a location check-in, on a popular location-based social networks (i.e., Foursquare). Such a check-in contains geographical (*i.e.,* address) and semantic (*i.e.,* venue category) information. In addition, the service provider extracts semantic mobility profiles across its users (*i.e.,* 'Places people like to go after. . . '), *etc.* The most relevant pieces of information are circled (in red). Source: [7].

To protect their privacy, users can obfuscate their location information, both at the geographical and semantic levels. For instance, a user can generalize[2] the semantic information of the venue by sharing, "restaurant" instead of "burger joint". The user can also generalize the geographical location of the venue by sharing the city instead of the exact address of the venue. By decreasing the likelihood that an adversary can track the location and activities of a user over time, location obfuscation provides users with better privacy guarantees. However, this could come at the cost of a reduction in users' perceived quality of service (i.e., utility) and possibly prevent users from achieving the objective they had in mind when they checked-in. It should be noted that utility is a general and subjective concept: It heavily depends on the users' context and perception (hence the terminology "perceived utility"). In this work, and more specifically in the user survey on which it is based, we define perceived utility as "the extent to which the initial [that is, before location obfuscation by generalization is applied] purpose of a check-in is still met [after location obfuscation is applied]". As our results are based on user data obtained through the survey, this definition should be retained as the formal definition of utility for our work, especially when interpreting the experimental results.

In general, it is difficult for users to estimate the privacy risks that stem from location sharing (this usually requires performing statistical inference [9]) and to make rational privacy-related decisions [15]. Moreover, it would be cumbersome for users to manually select the level of obfuscation that they want to apply to every single one of their check-ins. Therefore, *automatic* obfuscation mechanisms are needed (note that automatically generated privacy recommendations are valuable as well [16]). To balance privacy and utility, such mechanisms must be able to quantify the effect of obfuscation on both privacy and utility. Formal frameworks have been proposed to quantify location privacy, e.g., [9]. Note that privacy depends on the adversary being considered (other individuals, service providers, etc.) and on the user's perception (e.g., perceived sensitivity of different types of private data). Still, utility is often an overlooked aspect of the problem: Indeed, only a few studies address the utility loss due to location obfuscation for particular location-based services [17], [18], or the utility loss in a formal framework for finding the optimal balance between utility and privacy [19]. Despite these studies, there is no methodology for modeling and predicting the perceived utility loss that stems from the use of obfuscation mechanisms in location-based social networks for each individual check-in (for each individual user). In this paper we present such a methodology to design automatic and personalized location-privacy protection mechanisms.

The problem of predicting a user's perceived utility loss due to obfuscation is deeply intertwined with the problem of identifying *why* the user shares her location in the first place. In this paper, we propose to first infer the motivation of the user in sharing her location, and then to predict the utility implications of a privacy-protection mechanism on the user's experience with respect to that particular motivation.[3] Using this approach, we can determine which level of location obfuscation is acceptable to the user. For example, a user might only want to convey the message that she is performing a certain activity, such as "eating" in a given city, without revealing the exact type or address of the place where the activity is happening. In another example, consider a user who checks in to a restaurant in Hawaii; if her motivation is to invite some friends, then the full address of the venue is needed, but if she wants to let her friends know she is having a good time on vacation, then coarse grain information about the place, e.g., "restaurant in Hawaii", suffices.

We rely on machine-learning algorithms to find the right balance between the level of obfuscation and the utility requirements of each user. Given some features about a check-in (and the user's behavior), our algorithms predict user's motivation for this check-in and her perceived utility loss for each level of (geographical and semantic) location obfuscation. The result of our algorithms is a personalized utility function. We implement and test our methodology on the results of an online survey involving 77 Foursquare users (with 45 check-ins per user, hence a total of 3,465 check-ins). We predict the fine-grained purpose of the check-ins (among 13 pre-selected purposes) with a raw correct classification rate of 43% (63% for the coarse-grained purpose, among 4 pre-selected such coarse-grained purposes) and the effect of obfuscation on utility (on a scale from 1 to 5) with a mean prediction error of 0.52, significantly improving the performance of our model compared to the original version of this work [1].

Our survey's results also shed light on the effects of location obfuscation mechanisms on the perceived utility

2. In this work, we focus on the case of obfuscation by *generalization*. The case of obfuscation by *addition of fake information*, as proposed in the context of location privacy, is not considered in this work as it raises complex additional questions about utility. In [14], Wang et al. study the motivations behind *fake* check-ins.

3. Throughout the paper, we use the equivalent expressions *motivation behind* and *purpose of* check-ins interchangeably.

by users in location check-in applications. In particular, our results indicate that semantic obfuscation (e.g., reporting "restaurant" instead of "burger joint") has a significantly larger negative effect on the perceived utility, compared to geographic obfuscation (e.g., reporting the city instead of the full address). Beyond helping model the perceived utility, inferring the purposes of individual location check-ins can also be useful for creating new features on LBSNs. For example, users could be offered the "directions to the venue" feature for check-ins for which the purpose is that friends join, or be offered to share a group picture for check-ins for which the purpose is to inform about the people around them. More generally, the classification of the check-ins (w.r.t. their purposes) could be used to automatically adjust the way the check-in history is presented to the users.

In summary, our contributions are as follows:

1) We present the first, to the best of our knowledge, methodology for inferring the motivations behind users' location check-ins and their effect on users' perceived utility loss that is caused by different levels of location obfuscation (for both the semantic and geographical information).
2) We design a utility function that can be used as a building block for designing usable utility-aware location privacy-protection mechanisms. Such mechanisms could automatically choose the *optimal* obfuscation level that matches the users' preferences in terms of utility (or simply make suggestions and let the users choose).
3) We study the trade-off between utility and privacy in a location-based social network, namely Foursquare, based on the results of a survey of Foursquare users.

We make the following additional contributions, with respect to the original version of this work [1]: (1) We study in more detail the motivations behind location check-ins by looking at the secondary purposes that the users self-reported. (2) We analyze the differences between the check-ins that contain text messages and those that do not. (3) We propose and use a hierarchical model for the purposes which enables us to predict, with a higher accuracy, the coarse-grained purposes behind check-ins.; (4) We update our predictive model for the perceived utility loss; by replacing a simple regression model with a cost-sensitive multi-class classifier, we substantially improve the mean error of the prediction. Moreover, we release a sanitized version of our dataset, which we describe in detail in Section 7. Note that we also slightly improved the performance of the purpose classifier by considering a few additional relevant features, including the distance between the venue and the user's home.

The rest of the paper is organized as follows. After discussing the related work in Section 2, we present the methodology of our study in Section 3, which includes an online survey with Foursquare users, and the definition of the motivation and utility inference frameworks. Subsequently, we present quantitative results, by discussing both descriptive statistics and performance values of our motivation classifier and utility model in Sections 4 and 5, respectively. We describe the sanitized version of our dataset of self-reported purposes and perceived utility for Foursquare check-ins (our "utility dataset") in Section 7 (we release it to the community). We then discuss the limitations of our study, conclude the paper and give directions for future work in Section 8.

2 Related Work

From a high-level perspective, there are two broad categories of study on location-sharing behavior and privacy that are related to our work: (i) users' motivations for sharing location in online social networks, and (ii) location-obfuscation techniques and their effect on perceived utility.

2.1 Motivations behind Location Sharing

Recently, several works investigated users' motivations for disclosing their locations in online social networks. Patil et al. [3], [4] carried out two online user-studies ($N = 401$ and $N = 362$ participants, respectively) and studied the users' motivations for sharing locations on location-based social networks (in particular on Foursquare). The results show that users' main motivations include the desire to connect with their social circles and to project an interesting image of themselves. In particular, their motivations for sharing location information included the desire to tell friends that they like a place, to keep their social circle informed of where they are, to record their visits and to appear "cool" and interesting. As a consequence, the primary reason for "checking in" appears to be related more to attaining a higher-level objective, such as sharing a positive experience or appearing "cool", rather than to pointing to a specific geographical location. Similarly, results presented in [5], [6] also show that social connections and impression management play a cardinal role in users' location-sharing activities in Foursquare. Following these results, we adopt the motivation labels described in [3], [4] as the default options available to users for selecting the main purposes of their check-ins. In order to not restrict users to one of the predefined choices, we also offer them the option of entering a purpose that is not present in the predefined list. Cramer et al. [20] performed an in-depth qualitative study of users' motivations for checking in on Foursquare (e.g., reasons, context, audience), based on interviews ($N = 20$) and survey responses ($N = 47$). The main reasons for sharing location, which they extracted from their interview responses, match the motivation labels considered in this paper. One of their findings is that check-ins serve a *utilitarian* purpose (e.g., coordinate with friends), which shows the need for utility models (that we provide in this paper). The authors also investigate the importance of the audience of check-ins and the perception of a user's check-ins by her friends. Kim et al. [21] study, based on a survey with college students ($N = 255$), the relationship between privacy concerns and motivations behind location check-ins. Shajung et al. [22] study, based on a survey of Facebook users ($N = 523$), the relationship between personality and check-in behavior. Although related to our work, none of the aforementioned papers tackles the inference of the motivation behind check-ins and the design of (motivation-based) utility models for check-ins when using location

obfuscation techniques. In the same line as our work, the system proposed by Zhu et al. [23], for inferring the purpose of a trip (or the targeted activity at the destination), based on (semantic) crowdsourced data (i.e., Puget Sound Research Council survey data enriched with Foursquare venue data) fed to a SVM classifier. Rachuri et al. [24] propose a system for inferring a user's activity, using data from so-called soft sensors (e.g., app usage); the purpose of making such predictions is to assist users in writing the text associated with their location check-ins. For instance, if the system detects that a user is eating, it would automatically suggests a message of the form "Eating at..." when she starts a new check-in. They validate their approach on a real dataset collected through a field experiment ($N = 20$). Finally, on a related topic, Wang et al. [14] study the motivations behind *fake* check-ins, and show, using a survey of Facebook users ($N = 23$), that the main cause is external rewards provided by the platform (e.g., badges).

2.2 Location Obfuscation

Location privacy is a well-studied topic in mobile networks. Many location obfuscation mechanisms have been proposed, including reducing the granularity of the location (generalization), adding noise to the geographical location, adding fake location information, hiding location information, and changing identifiers [17], [25], [26], [27], [28].

Brush et al. [29] study the users' preferences and concerns for several such algorithms by showing the result of each of them to the users. Although the evaluation showed that the users understood the basic effects of the different algorithms, the authors highlighted a significant lack of awareness of long-term threats. A related effort by Tang et al. [30] presents the users with three different visualizations of their past shared locations; they study their effect on the end-user privacy. They show that, depending on the type of visualization, the users' self-reported attitudes diverged in terms of the people with whom they shared their locations.

There are also targeted studies on the usability of the proposed location obfuscation techniques for mobile applications [18], [31]. In particular, Micinski et al. [18] study the relationship between location obfuscation and application utility on the Android platform. By means of an Android tool, called CloakDroid, they show that providing applications with location information that is less precise does not substantially hinder their functionality. A more encompassing approach, taken by Henne et al. [31], enables Android users to specify different obfuscation algorithms for each Android application, including location truncation.

As users are not able to anticipate the privacy threats against themselves, caused by the information they share, there were several attempts to formalize the desirable location privacy requirements that obfuscation mechanisms should fulfill and the metrics to quantify them. Examples of such works are Krumm's [32], Decker's [33], and Duckham's [34]. In a follow-up of these works, Shokri et al. provide a framework [9] to quantify location privacy, and a game-theoretic methodology [19] to optimize location privacy while respecting users' utility requirements. Despite all the efforts to design obfuscation mechanisms and quantify their effect on users' location privacy, no methodology is proposed for quantitatively estimating the utility loss caused by different obfuscation mechanisms. Shokri [35] proposes to optimize utility under privacy constraints by taking a game-theoretical approach. The framework requires a cost function, which is precisely what we provide, in order to evaluate the utility loss caused by data perturbation, which is precisely what we provide. The few studies that include the utility aspects of location-obfuscation mechanisms reflect only the application dimension of it, for example, by measuring the fraction of restaurants that a user misses, the error of traffic information due to location perturbation [17], [18], the perceived loss of quality-of-service (quantified through user surveys) for location-based place finders [36], or the error made by a prediction service operating on location data (e.g., Yang et al. [37] use the error from a travel purpose-prediction mechanism [23] as a metric for the utility loss). Our work completes this line of studies by providing a methodology to design user-centric perceived utility functions for location check-ins in location-based online social networks.

3 Survey and Data Collection

In this work, we investigate (on a per-check-in basis) the effect of geographical and semantic location obfuscation (*i.e.,* generalization) on the perceived utility of (Foursquare) check-ins. In order to better understand users' behaviors and preferences when they check in at venues, we ran a user study in early 2014. The study consists of a personalized online survey, where participants are asked to provide additional information about their past check-ins on Foursquare. Foursquare is a very popular location-based mobile social network (unlike Facebook, users can only check-in from their mobile devices), whose primary feature is to check-in at venues: From the Foursquare mobile application or website, users can select a venue close to their current location (from the Foursquare database) and share their presence at this venue, possibly together with a text message and some pictures.[4] Each venue is associated with a street address and a semantic tag (from a predefined set of tags, organized as a tree).[5] Foursquare also provides incentives (e.g., badges, "mayorship", and rewards upon check-in) and gaming features (e.g., treasure hunts in which participants must check-in at specific venues).

In the survey, we ask the participants to state the purpose of some of their past Foursquare check-ins, as well as to specify to what extent the purpose of their check-ins would still be met if their check-ins were obfuscated at several levels (both geographical and semantic). Our findings are then used to evaluate an automated system that predicts the purpose and the extent to which such a purpose would still be met, if the original check-in were replaced by an obfuscated version of it.

In the following subsections, we discuss the details about the participants and the contents of the survey.

4. We chose Foursquare because of its popularity and because check-ins constitute its main feature. Moreover, its API enabled us to easily access all the information required to generate the survey.

5. https://developer.foursquare.com/categorytree. Last visited Dec. 2016.

3.1 Participants and Remuneration

To recruit participants, we made use of the Amazon Mechanical Turk (MTurk) platform; it enabled us to draw candidates from a pool of users with diverse backgrounds [38] and to limit the bias of the results towards academic and student behavior, inherent to on-campus surveys. We screened participants according to the following admission criteria: (i) aged between 18 and 80 years, (ii) with an active Foursquare account, (iii) with at least 75 check-ins over the last 24 months,[6] (iv) with at least 20 check-ins containing some text. Furthermore, to ensure a minimal level of diversity in the check-ins, we allowed only the participants who had checked-in at, at least, 15 different venues, stemming from at least 5 different venue types (with at least 2 different venues for each type). Note that we considered only venues that have both precise geographic and semantic information, and that have a non-negligible number of unique visitors. Moreover, we screened the MTurk participants according to their past performance on the platform: They had to have a minimum human intelligence task (HIT) approval rate of 95% and at least 100 past approved HITs. This was a preliminary step to preventing inexperienced and sloppy workers from participating in our survey.

Our survey is based on the participants' actual check-ins on Foursquare posted over the last 24 months (we collected through a specific application we developed), and it requires a significant amount of time to complete (30-45 minutes). To encourage the participants to participate in the survey and to temporarily grant us access to their Foursquare data, we rewarded them with a fixed amount of money (US $4.5 per HIT [39], [40]). At the end of the study, the average per-hour remuneration for the participants was US $8.50.

3.2 Online Survey

The survey, divided into two parts, was composed of a total of 68 questions. In the first part, participants replied to 18 questions pertaining to general demographics, as well as technology and location-sharing habits. The remaining 45 questions were constructed by using information collected from the users' own Foursquare check-ins.

Before beginning the survey, the participants were presented with a welcome page that indicated the scope and purpose of the study. After agreeing with the privacy and data-use policies[7], they were asked to log in to their Foursquare account and grant us access to their check-ins and friend lists. After this step, our application verified if the participants actually fulfilled the admission criteria and, if so, it allowed them to continue to the first (static) set of questions.

Following the first part, the participants were presented with the second (personalized) part of the survey, where they answered a set of questions for each of the 45 check-ins,

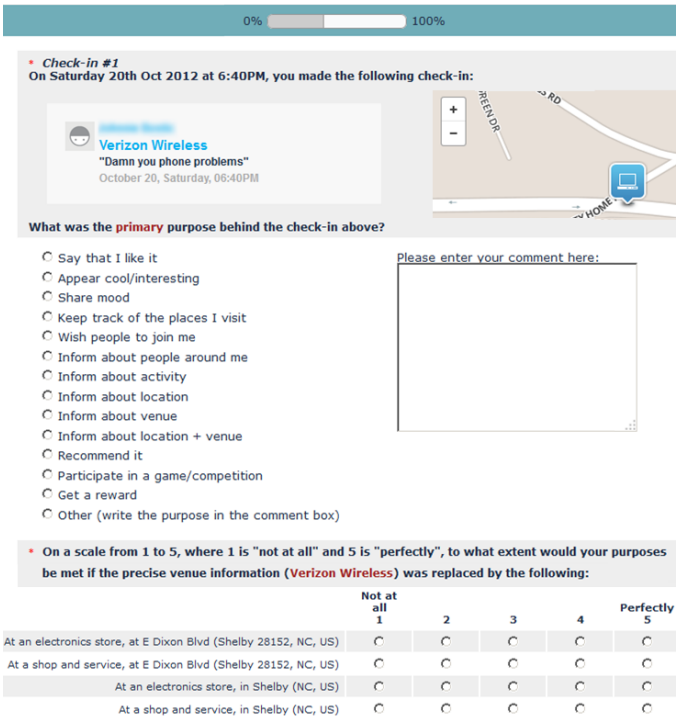

Figure 2. Screenshot of our online survey website. Participants are presented with some of their own past Foursquare check-ins and they are asked some questions about the purpose of their check-ins and the effect of (geographical and semantic) location obfuscation on their perceived utility. For privacy reasons, we blurred the name of the participant.

totaling 405 personalized questions. For each of their check-ins, the participants were presented with the time of the check-in, the venue (its name and its location displayed on a map), and the associated text message, if any (see Figure 2).[8]

These questions allowed participants to select one answer per question item, among a set of pre-defined choices. We asked participants to state (1) the primary and (optionally) secondary purpose of the check-in, (2) whether the text in the check-in is related to the location, (3) the extent to which the purpose of the check-in would still be met if it were replaced by a less detailed check-in (we had four different versions with varying levels of geographical and semantic obfuscation), (4) the most important detail in the check-in and (5) the most similar check-in in terms of purpose, among two other suggested check-ins present in the user's own questions. In particular, for (1) we allowed users to either select one among a set of 13 proposed choices (based on [3], [4] and our internal experiment) or to specify a different one in free-text. In the end, we only made use of (1) and (3).

We considered two levels of obfuscation (low and high), both at the geographical and the semantic levels. Geographic obfuscation reveals only some of the geographic information (the street number, street name, zip code, city, state, and country); semantic obfuscation reveals only an ancestor, in Foursquare's semantic hierarchy, of the semantic tag of the venue (in our dataset, semantic tags have 3 to 4

6. In order to prevent survey participants from artificially creating new check-ins only to match the admission criteria, only the check-ins made over the 24 months before the *starting date* of the survey were taken into account.

7. They approve a data retention and processing agreement, informing them that all data collected in our study is used solely for the purpose of our academic research project, and that we will not disclose or use it in any other way than what is explicitly mentioned.

8. Note that we did not include the pictures associated with the check-ins; in our dataset, only 6% of the check-ins contained pictures.

Table 1
Example of alternative check-ins with different combinations of geographical and semantic obfuscation levels.

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)

ancestors). The four combinations of obfuscation levels are defined as follows and are illustrated on a sample venue in Table 1:

1) *Low semantic obfuscation, Low geographical obfuscation (Ls-Lg)*: Instead of the full venue information, we show only the immediate ancestor in the semantic hierarchy of the venue, and we display only the street name/city/state/country (without the street number).
2) *High semantic, Low geographical (Hs-Lg)*: We show the second ancestor, and display the street name/city/state/country.
3) *Low semantic, High geographical (Ls-Hg)*: We show the immediate ancestor, and display the city/state/country.
4) *High semantic, High geographical (Hs-Hg)*: We show the second ancestor, and display the city/state/country.

Geographical obfuscation relies on the Google Geocoding API to convert the venue addresses to a structured format (street number, street name, zipcode, city, state, country), whereas semantic obfuscation relies on the tree structure of the set of tags provided by Foursquare. Table 1 shows an example of a check-in with the four alternatives, where a participant has to state, on a discrete 5-point scale (where 1 means "Not at all" and 5 means "Perfectly"), the extent to which her purpose would still be met if her original check-in were replaced by each of the alternative ones. As the results presented in this work are based on the survey data, this definition should be retained as the formal definition of perceived utility. It can be assumed that a users' perceived utility includes both the benefits she gets by making the check-in and, by transitivity [41], the benefits her friends get by reading her check-in (as considered in [42]). Note also that the utility loss caused by location obfuscation (and the purpose behind a check-in) might depend on the audience of the check-in. However, as Foursquare offers only two options for the audience of a check-in (public or friends), this aspect cannot be studied through this survey and is therefore left for future work. Figure 2 shows a lightened screenshot of our survey website for a sample check-in.

In order to detect and discard sloppy answers, we performed two tests: time analysis and purpose diversity. For both parts of the survey, we analyzed how long it took participants to complete them, and we discarded the participants whose timings were lower than twice the standard deviation around the mean time. Regarding the diversity in the stated purpose, we retained participants who chose at least two distinct purposes at least twice in their answers. To avoid wasting participants' time, we did not include "dummy" questions in the survey, as our previous experience showed they were answered correctly, even by the participants who provided sloppy answers.

3.3 Statistics about the Participants

After filtering out participants who did not meet the admission criteria, we obtained a total of 77 valid questionnaires. The average age of the respondents were 29$\pm$6 years, where the oldest and youngest participants were 50 and 19, respectively. 57% were female, and the participants were almost all based in the US (96%). The other participants came from Canada (1), Norway (1) and Israel (1). Only 14% of them were students, whereas the rest of them listed occupational sectors such as education (12%), medical (8%), and arts and entertainment (8%). Only 7% of participants stated that they were unemployed.

When asked about technology usage, all respondents reported to have been using social networks for more than 2 years, with 67% of them connecting once per week or more often. With respect to privacy on the Internet, on average the participants were mildly concerned (average score of 2.9 on a 5-point scale, where 1 means "not at all" and 5 means "very much"). A similar result was observed when we asked about their level of comfort when other people "tag" them at different locations (score of 2.2 on a 5-point scale, where 1 means "not at all" and 5 means "very comfortable").

3.4 Purposes of Check-ins

In the second part of the survey, participants were asked to provide the main purpose for each of their 45 check-ins.

Overall, the 13 purposes that participants could select from were sufficient to explain, as a primary purpose, 99% of all 3465 check-ins. Figure 3 shows the distribution over the participants' primary and secondary purposes for their check-ins. We can observe that, among the top four primary purposes (which account for 63% of all check-ins), there are only those that are either related to higher-level social goals (such as informing about their current activity or mood[9]) or to personal record-keeping purposes, which corroborates the results obtained by [3], [4]. The purpose of informing about the actual location was selected only for less than 9% of the check-ins. As for the secondary purposes, we observe a similar trend (note that in 28% of the cases, the survey participants did not specify a secondary purpose).

9. Recent research has shown that mood is correlated with location and context in general [43].

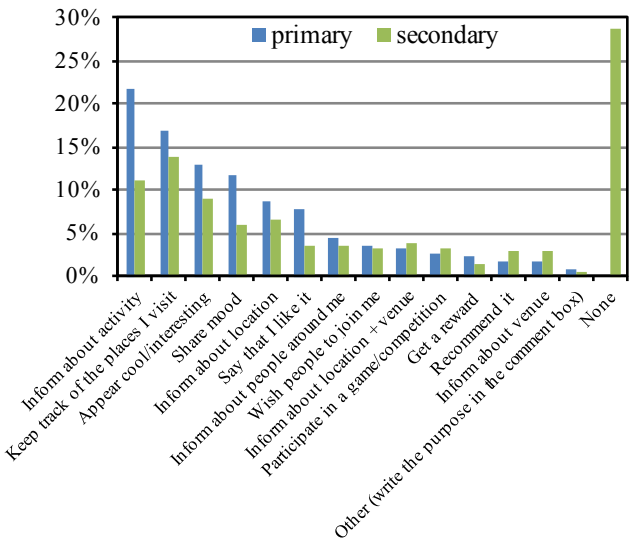

Figure 3. Proportion of primary (left bars) and secondary (right bars) purposes for the users' check-ins. The top four primary purposes, which account for 63% of the total, represent only high-level social and personal goals. Informing about the actual location is only the 5th most frequent purpose, selected in less than 9% of the cases. Some check-ins do not have a secondary purpose (28%).

In spite of such a large difference between the first and second group of purposes, we are aware of only one major social network (Facebook) that allows users to share their mood in a structured way, in addition to the actual post or check-in. Other providers, such as Twitter or Foursquare, do not yet provide this possibility; they rely on users to express their mood in an unstructured way in their messages.

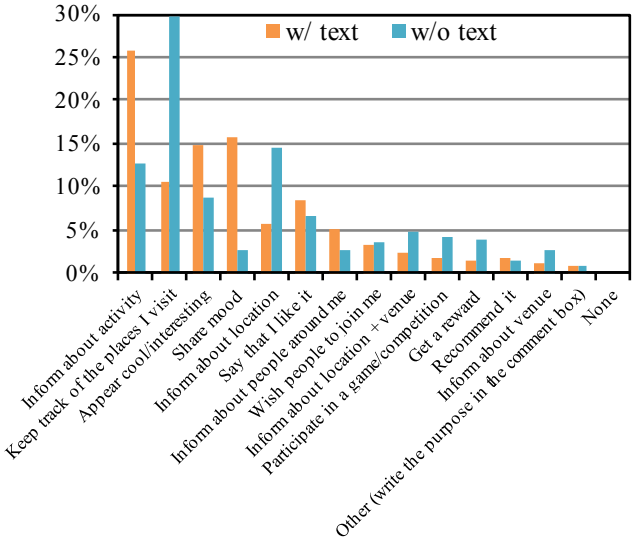

Figure 4. Proportions of primary purposes for the users' check-ins, with (left bars) and without (right bars) text attached to them.

We also compare the distributions of primary purposes, for check-ins with and without text, depicted in Figure 4.[10]

10. This is based on the fact that the `has_text` feature is one of the most influential, according to our preliminary analysis based on the information gain metric; we explain this in more detail in Section 4.2.

Check-ins with text account for 68% of our dataset. We observe a striking difference between the two distributions, which demonstrate there exists two distinct use-cases for check-ins. For check-ins without text, the most frequent purpose, by far, is "Keep track of the places I visit": People use such check-ins to keep a record of the venues they are at and this goal can be met without the need to attach a text message. For check-ins with text, however, higher-level social purposes, such as "share mood", are more represented: Reporting solely the venue a user is at is clearly not enough to convey her current mood; a text message is therefore needed to achieve this purpose.

Finally, in order to analyze the purposes of the check-ins at a higher level (see Section 4.2.2), we propose a purpose hierarchy by clustering all the 13 fine-grained purpose labels into 4 coarse-grained purpose labels. Figure 5 depicts the proposed hierarchy. The corresponding distribution of coarse-grained purposes (based solely on reported primary purposes) on the entire dataset is as follows: informative (40%), utilitarian (22%), personal (33%), gaming (5%).

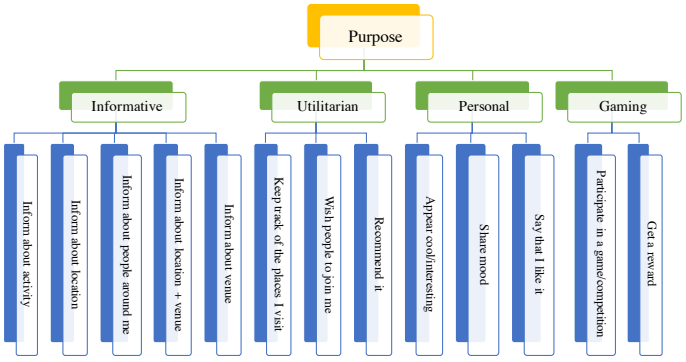

Figure 5. Proposed hierarchy for check-in purposes. The 13 fine-grained purpose labels are clustered into 4 different coarse-grained purpose labels.

3.5 Utility of Check-ins vs. Obfuscation Levels

Given the aforementioned findings, hereafter we investigate the effect of the reduction of details in a check-in on its perceived utility for the user. We define "utility" as the extent to which the purpose of a check-in is still met after an obfuscation function (which removes some information about the check-in, as shown in Table 1) is applied. In our survey, participants selected the utility value on a discrete 5-point scale, where 1 means "Not at all" and 5 means "Perfectly".

First, we study the relationship between obfuscation and utility in general, where we do not distinguish between the different purposes of the check-ins. Second, we perform this analysis on a per-purpose level, showing that the purpose mediates the effects of obfuscation on the utility. These findings constitute the basis for the development of our purpose-inference framework and our utility-obfuscation model.

3.5.1 Utility vs. Obfuscation (in General)

In order to study the general relationship between utility and obfuscation, we group the check-ins according to

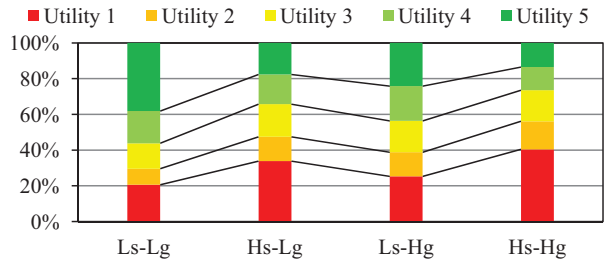

Figure 6. Proportion of check-ins with their perceived utility, for different levels of geographical and semantic obfuscation. A utility of 1 means that the purpose of the check-in is not met at all after the obfuscation, whereas a utility of 5 means that the purpose is met perfectly after the obfuscation. Perceived utility decreases with the level of obfuscation; semantic obfuscation has a stronger (negative) effect on utility.

the four combinations of obfuscation levels, described in the section "Survey and Data Collection", i.e., (Ls-Lg),(Hs-Lg),(Ls-Hg),(Hs-Hg). The results are depicted in Figure 6.[11]

We observe that even with the lowest obfuscation level (Ls-Lg), 38% of all check-ins would still keep a maximum utility, whereas, for 21% of them, the utility would be severely affected. When the level of semantic obfuscation increases (Hs-Lg), there is a sharp increase (+70%) of the check-ins that would lose all utility, and a significant decrease (-50%) of those that have maximal utility. Hence, semantic obfuscation has a sharp negative effect on the utility of check-ins. However, in the scenario where it is the geographical obfuscation that increases instead of the semantic (Ls-Hg), the results show that there is only a moderate increase (+25%) of check-ins with the lowest utility, compared to the base case Ls-Lg, and a moderate (-37%) decrease of the check-ins that would still keep a maximum utility. Therefore, compared to the geographical obfuscation, our results indicate that the semantic obfuscation has a greater negative effect on utility.

3.5.2 Utility vs. Obfuscation (Given the Purpose)

Figure 7 shows the participants' utility scores for check-ins, grouped according to their purpose: "inform about activity" (Figure 7a), "appear cool/interesting" (Figure 7b), and "wish people to join" (Figure 7c).

For the check-ins with the purpose of informing others about the user's activity (which is the most popular purpose with 22% of total check-ins), we observe an even stronger effect of semantic obfuscation on the utility, compared to geographical obfuscation. In particular, compared to the Ls-Lg scenario, the lowest utility score increases from 19% to 40% (+111%), when the semantic obfuscation is increased; however, by increasing the geographical obfuscation, the same utility score increases only from 19% to 21% (+11%). A similar message is conveyed by the sharp decrease of the highest utility from 39% to 7% (-83%) for the high semantic obfuscation, as compared to only a -42% for high geographical.[12]

11. The differences among the averages of the four obfuscation levels are statistically significant, both pairwise and globally (a χ^2 test of homogeneity gives a p-value: $p < .01$). The p-value is computed from the data and quantifies the significance of the observed difference.

12. $p < .01$

For check-ins with the purpose of appearing cool/interesting (Figure 7b), the utility scores exhibit lower variations as compared to Figure 7a and more in accordance with the general motivation-utility results shown in Figure 6.[13] An interesting result is shown by Figure 7c, where the purpose of the check-ins is "wish people to join". In this case, we do not observe any significant differences between semantic and geographical obfuscation on the utility scores; in fact, the only statistically significant one is between Ls-Lg vs. Hs-Hg ($p < .05$). Hence, as expected, it seems that any kind of strong obfuscation has a largely negative impact on the utility of this kind of check-ins. Nevertheless, the presence of 25% of obfuscated check-ins with a maximum utility score might suggest that, for these users, wishing people to join them could be interpreted as a wish for other people to get in touch with the user, in order to obtain more detailed information about his precise location. Then, the user could engage with other people in a more interactive way, through other means (phone call and/or messages). Further investigation of specific cases is an interesting objective that we intend to pursue as future work.

The results presented so far show that the purpose of a check-in can indeed mediate the effect of different types of obfuscation techniques (semantic and geographical) on the perceived utility. Using our findings, in the two following sections, we describe and evaluate (on the data collected in our survey) an automated purpose-based utility model for location check-ins on Foursquare. Our solution is split into two blocks (Figure 8): First, we present a framework to infer the purpose of check-ins, based on a number of features extracted from the check-ins (e.g., location, semantic and textual information). Second, we present a utility model that uses, among other features, the (inferred) purposes of check-ins to predict the utility loss caused by the use of different obfuscation techniques.

4 Check-in Purpose Inference

A location check-in usually consists of two parts: The structured venue information (geographic coordinates and semantic hierarchy) and an (optional) unstructured text input by the user. In our work, we derive meaningful features for both parts by taking advantage of techniques from Natural Language Processing (NLP) and by crafting features specific to location-sharing on social networks.

4.1 From Check-ins to Features

]Using the check-in information, we extract features that describe the check-ins and that might encode information about the user's motivation for checking-in, having the Foursquare context in mind (e.g., distance to home, mood extracted from the text). The different features are combined in a single feature vector that will be fed to the machine-learning algorithm (i.e., a classifier), in order to derive the most likely purpose for each check-in. Hereafter we describe all the different components of the feature vector. The list of the features included in our sanitized dataset (described in Section 7), which constitutes a subset of all the features we

13. All differences are statistically significant at $p < .01$, except for Hs-Lg vs. Hs-Hg for which $p < .05$.

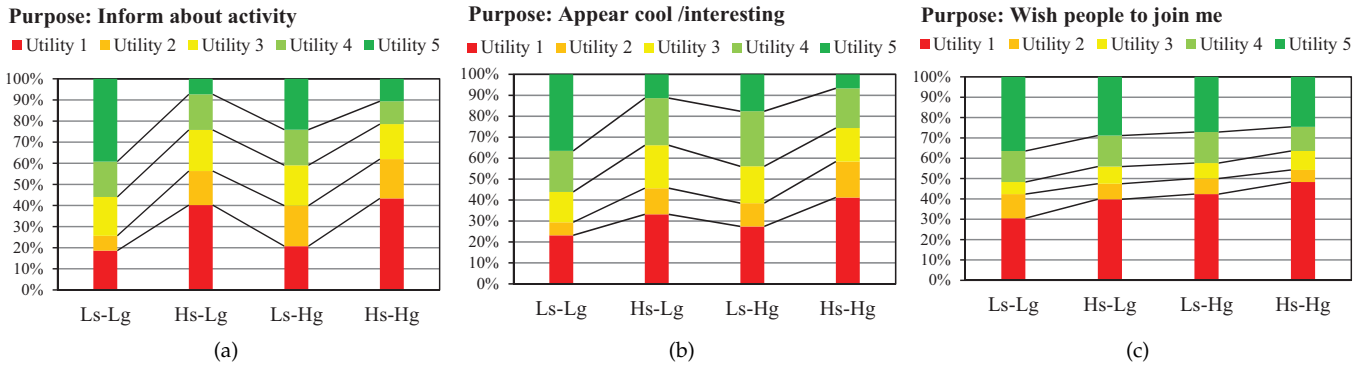

Figure 7. Proportion of check-ins with their perceived utility, for different levels of geographical and semantic obfuscation, according to their purpose.

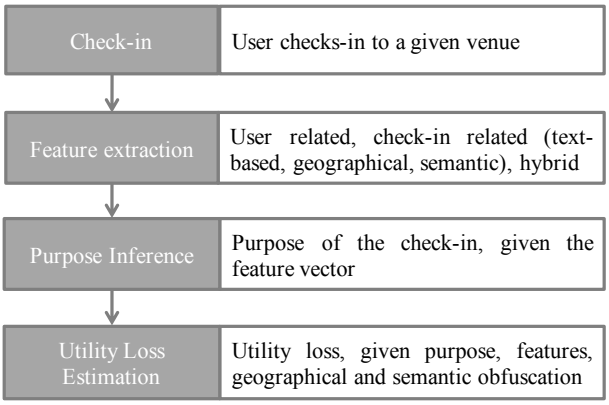

Figure 8. Workflow of the utility model framework, including the purpose inference stage.

used in our evaluation, is given in Table 6 (p. 17). Note that, for practical reasons, we consider only features that can be extracted from the information that would be available to an automatic tool running on the users' devices. Regarding the feature selection, it is performed automatically upon learning the model. Standard techniques include information gain-based feature selection (InfoGain) and correlation-based (CSFSubset) [44] feature subset selection. The former measures how each feature contributes in decreasing the overall entropy, and the latter measures the individual predictive ability of each feature, along with the degree of redundancy between them. Note that some classifiers (e.g., Random Forests) perform this step internally upon training.

4.1.1 Structured Venue and User Features

By using the Foursquare API, we access the following data about each check-in: venue name and type, number of check-ins at a venue, and complete address. Moreover, we extract the user's age, total number of check-ins, occupation and gender. Such features enable the classifier to exploit the fact that users with similar demographic attributes might exhibit similar behaviors. Finally, in order to enable the classifier to detect similarities between venues based on the position of their semantic tags in the tag tree, we include the ancestors, in the semantic-tag hierarchy, of the venues' tags: Even though an Italian restaurant and an American restaurant have different semantic tags, they are somewhat similar; this can be deduced from the fact that they share the same parent tags, namely "restaurant".

4.1.2 Unstructured Text Features

Using prior studies in the analysis of short texts, we extract the following high-level text-related features from each check-in: the emotion (such as joy or anger) [45] and the sentiment (positive or negative) [46]. These features are determined from other low-level features such as n-grams, punctuation marks, emoticons, capitals, key words and character repetitions. We used the Python NLP toolkit (NLTK 3.0) for the extraction of the low-level textual features[14], and we used a Naive Bayes classifier (trained on relevant short-texts [45], [47]) in order to extract the high-level ones. Such features can help us infer the purposes of check-ins; typically, it is less likely that the purpose of a check-in is "say that I like it" or "recommend it" if the emotion extracted from the associated text is "anger" and the sentiment is "negative". Several other pieces of work focus on the extraction of sentiment at the post/check-in level [48], [49], [50]. We also include some features that capture the presence of specific keywords (including punctuation and smileys) in the text associated with the check-ins. For instance, we capture whether the word "yummy" appears in the text. Such a feature typically enables the classifier to identify check-ins with purpose "Say I like it" (for restaurants).

4.1.3 Hybrid Features

To capture the correlation that might exist between the users' text and the venue information, we compute the longest common substring and, afterwards, the Levenshtein distance [51] between that substring and each field related to the venue. For instance, we determine whether the name and the city of the venue appears in the check-in text.

We also include some hybrid features computed from the user and venue attributes. For instance, for a given check-in, we compute the distance between the venue location and the user's home location. In addition to the raw distance, we include binary features that indicate whether the venue location and the user's home are in the same city/state/country. The rationale behind the inclusion of such features is that users might exhibit different behaviors depending on how far they are from home.

14. Available from https://www.nltk.org/. Last visited: Dec. 2016.

4.2 Inferring Purposes with Machine Learning

After we generate the feature vector for each check-in, we use it in a multi-class classifier to determine the most likely purpose of the corresponding check-in. We train the classifier on the data of all the users (10-fold cross-validation on the entire dataset). However, we retain the userID as one of the features. By doing so, we can benefit from a larger training set (compared to the case where the classifier used to infer the purpose of a user's check-in is trained only on the data of this particular user) while still being able to distinguish (through the userID feature) the check-ins made by the user who made the check-in which purpose is being inferred. This trade-off is discussed and analyzed in [40], [52]. Figure 8 shows the workflow of the entire inference process. We experimented with a Random Forest classifier [53],[15] using up to 100 trees of up to 10 features). Our results are obtained using the well-established WEKA toolkit [54], based on 10-fold cross validation. In order to avoid overfitting, we keep the default parameters used in the implementation of the various classifiers provided in WEKA. The source used to build the utility model from the survey data is provided with the sanitized utility dataset (described in Section 7). We use the data obtained through our survey as ground-truth to train the classifier (i.e., supervised learning) and to validate the results. We evaluate the performance of our prediction framework in inferring the fine-grained purpose of a check-in (among 13 labels, as described in Section 3.4) and its coarse-grained purpose (among 4 labels). We use the correct classification rate (CCR) as a performance metric and compare the performance of the considered classifier to that of a ZeroR classifier; this classifier always returns the most frequent label observed in the training set. The ZeroR classifier is the optimal classifier when no features are available. Using ZeroR as a baseline enables us to assess the benefits (and the potential) of using contextual features to predict the purposes of the check-ins.

For a first experiment, we determine the most influential features based on the information gain metric. The top features are, the (pseudo) user identifiers (such as user ID and hometown; this shows that check-in behaviors are highly personal), the venue type and location (note that the venue location might act as a pseudo identifier as our survey participants live in different cities), whether the check-in contains text, and whether the check-in contains a badge.

4.2.1 Inferring fine-grained purposes (13 labels)

Table 2 shows the performance of our purpose inference classifier (Random Forest), in the form of a confusion matrix, for the fine-grained purposes of the check-ins. These results are obtained from all the check-ins for which the participants specified a purpose (3435 in total). The cell at the intersection of row (a) and column (b) shows the number of check-ins with purpose (a) that are classified as purpose (b). The diagonal cells thus correspond to the correctly classified check-ins. For a global performance metric, we use the Correct Classification Rate (CCR): the proportion of check-ins for which the inferred purpose matches the actual one (i.e., the sum of the diagonal cells, normalized by the total number of check-ins). We obtain a CCR of 43%;[16] this has to be compared to the performance of a classifier that does not have access to any check-in information. When no information is available, the optimal classification consists in assigning the most frequent label to all instances (here, (c) "Inform about activity"), namely a ZeroR classifier. In this case, the CCR is the proportion of instances of the most frequent class (in the training set), that is 22% in our dataset. We use this as a baseline. Therefore, by using our features, the CCR is almost two times higher than the baseline. The relatively high number of possible purposes (i.e., 13) should be taken into account when interpreting the performance of the classifier: considering the number of labels, the CCR is relatively high, and constitutes a significant improvement compared to a feature-less classifier (i.e., ZeroR, our baseline). As explained in the next section, the CCR increases significantly when considering fewer (coarse-grained) purpose labels. The results are comparable to the those obtained in a similar setting [24]; we discuss ways to improve the performance in Section 6. Note that misclassifications have different levels of severity (classifying a check-in with purpose "Recommend it" as "Say I like it" can be considered closer, in terms of similarity, than classifying it as "Get a reward"). We relax the notion of correct classification rate to include the proportion of check-ins for which the inferred purpose is the self-reported *primary or secondary* purpose. In this case, the CCR increases to 48%.

We also look at the precision and the recall for each label (i.e., purpose). The precision for purpose (a) is defined as the number of check-ins with purpose (a) that are classified as purpose (a), normalized by the total number of check-ins classified as purpose (a), i.e., the diagonal cell divided by the sum of the cells of the column. The recall for purpose (a) is the number of check-ins with purpose (a) that are classified as purpose (a), normalized by the total number of check-ins with purpose (a), i.e., the diagonal cell divided by the sum of the cells of the row. Note that the recall corresponds to the correct classification rate within a class. High values of the precision and of the recall denote good performances of the classifier.

It can be observed that for the three most frequent purposes (i.e., (c) "Inform about activity", (j) "Keep track of the place I visit", and (d) "Appear cool/interesting), which cover more than half of the check-ins, the precision and the recall are significantly higher than the baseline, i.e., greater than 40%. The classifier performs best with check-ins with purpose (l) "Participate in a game"; this is probably due to the fact that such check-ins are specific to certain types of venues and that the text messages are automatically generated, hence easier to identify (the same applies to purpose (k) "Get a reward"). The classifier performs worse for check-ins with purpose (g) "Inform about venue"; this is probably because this purpose is quite generic, and because the proportion of such check-ins is too low to efficiently learn meaningful patterns while training.

Next, we consider the sorted lists of purposes returned

15. We experimented with various classifiers implemented in WEKA, with their default parameters. We report the results obtained with a Random Forest classifier as it is fast and gives, on our dataset, good prediction performance. Optimizing the performance through advanced models and parameters is out of the scope of this work.

16. We obtain 45% with a Support Vector Machine (SVM) classifier, 40% with a k-Nearest Neighbors (kNN) classifier.

Table 2
Confusion matrix for the 13-label purpose classifier, with the per-label precision and recall. The baseline is obtained by always assigning the most frequent label in our dataset (i.e., (c) "Inform about activity") to all the check-ins. Note that the CCR corresponding to the confusion matrix shown below does not match the average CCR reported in the text: the first is obtained from a single experiment, whereas the latter is aggregated over multiple experiments (with different seeds).

↓ Classified as →		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	Total (%)	Prec.	Rec.
Inform about location	(a)	101	1	97	13	10	17	3	8	10	25	1	1	10	297 (9%)	43%	34%
Recommend it	(b)	4	1	19	5	2	13	0	2	1	7	8	0	1	63 (2%)	5%	2%
Inform about activity	(c)	38	1	462	45	13	62	7	30	7	76	1	7	3	752 (22%)	46%	61%
Appear cool/interesting	(d)	9	0	79	171	17	53	1	24	11	71	5	4	4	449 (13%)	40%	38%
Inform about people around	(e)	8	1	45	16	34	17	3	6	8	12	0	1	3	154 (4%)	30%	22%
Share mood	(f)	17	3	83	58	9	166	2	22	4	36	1	2	1	404 (12%)	38%	41%
Inform about venue	(g)	4	0	22	2	4	6	1	3	1	16	0	0	3	62 (2%)	4%	2%
Say that I like it	(h)	7	2	62	33	9	33	1	71	1	44	5	2	3	273 (8%)	35%	26%
Wish people to join me	(i)	11	2	15	14	4	8	3	6	37	13	0	3	2	118 (3%)	39%	31%
Keep track of the places I visit	(j)	13	2	74	50	6	46	4	22	12	337	4	7	4	581 (17%)	51%	58%
Get a reward	(k)	4	5	7	6	0	3	0	1	0	13	39	4	0	82 (2%)	57%	48%
Participate in a game	(l)	4	0	7	7	2	3	0	2	1	6	5	54	0	91 (3%)	64%	59%
Inform about location + venue	(m)	13	1	32	10	5	7	0	4	3	8	0	0	26	109 (3%)	43%	24%

by the classifier (instead of looking at only the first purpose returned) and we look at the position (or rank) of the actual purpose of the check-ins in this list. Figure 9 shows the histogram and the cumulative distribution function of the rank. It can be observed that, in 61% of the cases, the actual purpose appears in the first two elements of the sorted list, and for 80% of the cases it appears in the first four elements. This implies, if users were to manually select the purpose of their check-ins from a sorted drop-down list, for 80% of the cases the output of the classifier would reduce the user burden (hence increase usability), as they would find the true purpose in the first four elements of the list.[17] In the baseline scenario, where a (feature-less) classifier simply returns the list of purposes sorted by decreasing frequencies, this numbers would drop to 39% (i.e., 22+17) and 64% (i.e., 22+17+13+12).

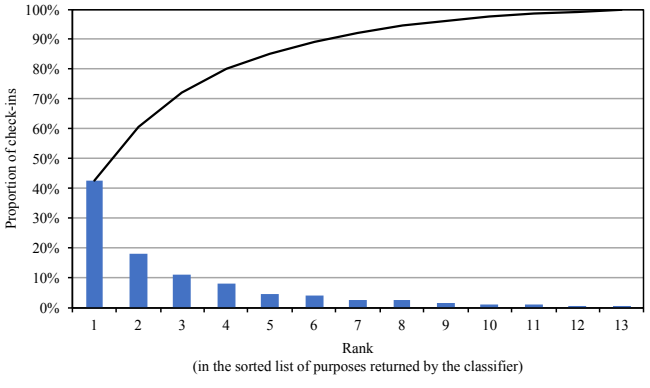

Figure 9. Rank of the actual purposes of the check-ins in the sorted list of purposes returned by the classifier.

Finally, we compare the raw performance of the purpose classifier for check-ins with and without text. Note that the experimental setup (Random Forest classifier, training and testing sets determined by 10-fold cross-validation, etc.) is the same as in the previous experiments; only the presentation differs, because we compute the CCRs separately for

17. Rachuri et al. [24] use the same methodology to evaluate their activity-prediction framework.

Table 3
Confusion matrix for the 4-label coarse-grained purpose classifier, with the per-label precision and recall.

↓ Classified as →		(A)	(B)	(C)	(D)	Total (%)	Prec.	Rec.
Informative	(A)	1024	101	236	13	1374 (40%)	66%	75%
Utilitarian	(B)	207	360	177	18	762 (22%)	60%	47%
Personal	(C)	288	119	702	17	1126 (33%)	61%	62%
Gaming	(D)	25	21	28	99	173 (5%)	67%	57%

the check-ins with and without text. For check-ins with a text message attached, we obtain a CCR of 41% (for a CCR of 26% for the baseline, namely a ZeroR classifier that returns the "Inform about activity" purpose label). For check-ins without text, we obtain a substantially higher CCR of 52% (for a CCR of 13% for the baseline). A possible explanation is that, although check-ins with text contain more information (which can be exploited through carefully selected, NLP-based, text-related features), check-ins without text are, by nature, simpler (this can be observed from the baseline CCRs already); it is therefore "easier" to infer the purpose behind check-ins without text.

4.2.2 *Inferring coarse-grained purposes (4 labels)*

Table 3 shows the performance of our purpose inference classifier, for the coarse-grained purposes based on the hierarchy presented in Section 3.4. The table has the same format as the table presented in the previous section and was generated in the same experimental setup. It can be observed that, when we cluster the 13 different fine-grained purposes into 4 coarse-grained purposes, the correct classification rates increases up to 63% (63% for check-ins with text and 65% for check-ins without text) for a baseline CCR of 40% (41% for check-ins with text and 38% for check-ins without text). Interestingly, even for the coarse-grained purpose "Gaming", which is not frequently reported in the dataset with only 5% of the check-ins, the precision is reasonable; this can be explained by the fact that such check-ins are in fact very specific.

Table 4
Summary of the results: performance, in terms of CCR, of the classifiers and their corresponding baselines.

		classifier	baseline
Fine-grained purpose (13 labels)	all	43%	22%
	w/ text	41%	26%
	w/o text	52%	13%
Coarse-grained purpose (4 labels)	all	63%	40%
	w/ text	63%	41%
	w/o text	65%	38%

Table 4 summarizes the different experimental results discussed in this section.

4.3 Discussion

The proposed system enables us to predict the purpose behind a check-in, based on a number of features extracted from it. The system is based on a supervised-learning approach, i.e., it learns from data that is manually labeled by the user. User check-in behaviors vary across users and vary

over time [20], [55], [56], therefore the system must cope with these variations. Regarding the variations across users, as explained above, for each check-in to be classified the userID feature enables the classifier to distinguish between the check-ins in the training set that were made by the same user and those made by a different user. Regarding the temporal variations for a given user, a possible solution (discussed in [40], [52]) is to assign higher weights (during training) to recent data in the training set or to simply remove old data from the training set and to re-train the model regularly.

5 Predictive Utility Model

In the previous section we show that a large proportion of predicted motivation labels are correct. This suggests a potential for exploiting automated methods for the inference of users' purposes for checking in on location-based social networks. More importantly, we now look at the perceived utility loss when (geographic and semantic) location obfuscation mechanisms are used; such models are crucial to designing utility-aware privacy-protection mechanisms.

In this section, we study the relationship between the purpose of a check-in and the loss of perceived utility, in the case where some of the details about it are obfuscated or not revealed. Ultimately, our goal is to define a predictive model of utility of a check-in, given the purpose (actual or inferred) of the check-in, the level of (semantic and geographical) obfuscation and characteristics of the venue and of the user (i.e., the features presented in the previous section).

5.1 Inferring Perceived Utility

We focus on the core technique for inferring the perceived utility. A more in-depth analysis of the factors causing the utility loss is available in the original version of this article [1]; this analysis is based on linear and non-linear regression models, which provides only modest accuracy. In this version, we propose a substantially more accurate prediction model.

As for the purpose inference, we rely on a multi-class classifier to decide between 5 different utility labels, specifically a perceived utility of "1", "2", "3", "4", or "5". However, unlike in the purpose inference case, not all prediction errors are equal: Predicting a utility of "1" when the actual perceived utility is "3" is worse than predicting "2". Therefore, we rely on a cost-sensitive classifier and define a penalty-matrix; by doing so, when training, the classifier optimizes the total incorrect classification penalty, instead of just the total number of incorrect classifications. Because we want to minimize the prediction error, we simply set the coefficients of the penalty matrix to the absolute value of the difference between the predicted utility and the actual utility, as shown in Table 5. As for the purpose inference, we use the userID as one of the features and train the classifier on the data of all the users, using 10-fold cross validation. The workflow is the same as for purpose inference (i.e., feature extraction, training and prediction–no post-processing). In fact, the features used in the utility inference are the same as for the purpose inference plus the obfuscation level and possibly the predicted purpose.

Table 5
Penalty matrix for the cost-sensitive classifier used for predicting perceived utility. The coefficients are set to the absolute value of the difference between the predicted utility and the actual utility (i.e., reported by the survey participants).

↓ classified as →	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

Figure 10 shows the distribution of errors for the prediction of the perceived utility. The results are based on a standard 10-fold cross-validation, with a cost-sensitive decision tree classifier (i.e., J48 [57]), configured with the penalty matrix depicted in Table 5. In a nutshell, J48 is a specific type of decision tree, in which at each node of the tree, a test is performed on a given feature and the leaves contain labels. It can be observed that the correct classification rate (i.e., which corresponds to the cases where the prediction error is zero) is 65%. The corresponding mean prediction error is 0.52 (for a baseline error of 1.9); the error should be interpreted with respect to the size of the utility range, i.e., $[1, 5]$. We believe that the error is sufficiently low to enable the design and implementation of efficient utility-aware privacy protection mechanisms. The proportion of cases where the predictor overestimates the perceived utility is roughly the same as the proportion where the predictor underestimates it. Note that this could be biased by introducing asymmetry in the penalty matrix, that is, by varying the relative cost of overestimation (penalties over the diagonal) vs. underestimation (penalties under the diagonal).

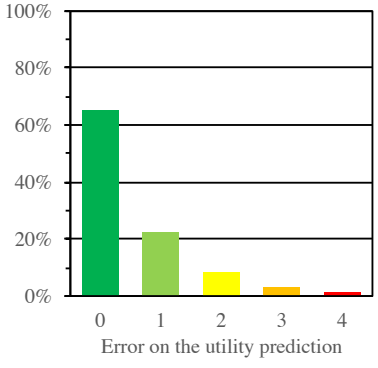

Figure 10. Distribution of the absolute error for the utility prediction.

5.2 Privacy/Utility Trade-Off

In our study, one straightforward way to take the privacy of check-ins into account is to quantify it through the different obfuscation levels. In particular, one can assume that the lowest level of privacy for a user is achieved when no information about her check-in is obfuscated; then, a slightly higher privacy level is achieved when low obfuscation is used on both the semantic and geographic levels (Ls-Lg). Then, an even higher privacy level is reached when either semantic or geographic levels are high (Ls-Hg or Hs-Lg);

finally, the highest level of privacy is achieved by the highest level of obfuscation on both the semantic and geographic levels (Hs-Hg). 3Note that the definition, and thus the quantification, of privacy depends on the considered adversary (i.e., the entity that has access to the data), typically other users (social contacts or not) and/or service providers (main or third-party). Therefore, in order to properly explore the privacy/utility trade-off, the adversary must be specified.

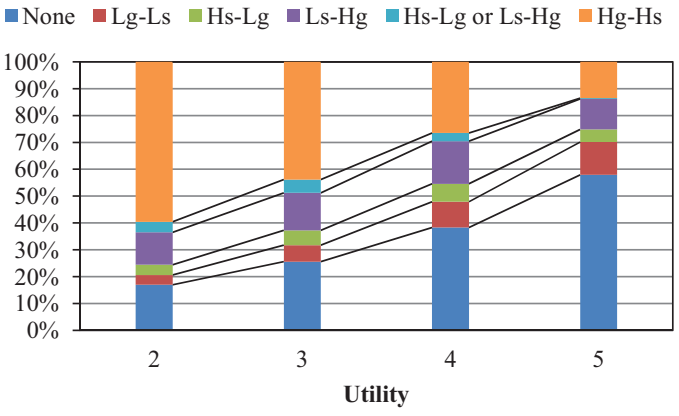

Figure 11. Proportion of check-ins that can be obfuscated to the highest level among the four semantic and geographical combinations, for a given utility value in the interval $\{2, \ldots, 5\}$. If no combination of obfuscation meets the utility value, the highest obfuscation combination is set to "None", i.e., we keep the full details of the check-in.

For each utility value in $\{2, \ldots, 5\}$, Figure 11 shows the proportion of check-ins, with respect to the highest obfuscation level that meets it. As the obfuscation levels Ls-Hg and Hs-Lg are not directly comparable, we distinguish between the cases where (1) Ls-Hg meets the utility threshold and Hs-Lg does not, (2) Hs-Lg meets the threshold and Ls-Hg does not, and (3) both do (that we denote by Ls-Hg or Hs-Lg, as any of the two levels can be used). If no obfuscation levels meet the utility threshold, the highest obfuscation level is set to "None". For example, a check-in with the following utility ratings (Ls-Lg: 3, Hs-Lg: 2, Ls-Hg: 3, Hs-Hg: 1) and a utility value of 5 cannot be obfuscated (hence its category is "None"), for a value of 3 it is "Ls-Hg", for a value of 2 it is "Ls-Hg or Hs-Lg", and for a value of 1 it is "Hs-Hg".

From the figure, we observe that even for very conservative users (who set the utility threshold to 5), 42% of their check-ins can still be obfuscated, and 13% of their check-ins can be obfuscated at the highest level (Hs-Hg). It is interesting to note that, for a relatively high utility value of 4, more than 60% of the check-ins can still be obfuscated, including 26% at the highest level. These findings are of great importance for service providers because they show that it is possible to find a balance between privacy and utility in location-based social networks; in fact, a large majority of check-ins can be obfuscated without incurring a significant loss of utility, which in turn enables social network providers to put privacy in the design of their systems with a negligible effect on their usability. For example, the utility values could be used to select the default obfuscation levels (semantic and geographic) for a given check-in, and enable users to change it in case it does not meet her utility preferences.

Furthermore, as the proposed mechanism can be executed entirely on the users' own device (in terms of purpose inference and obfuscation levels), there is no need for the service provider to store additional user information. This, in turn, provides an additional incentive for users to adopt it. In terms of execution time of the purpose inference and estimation parts, the users are not required to train the purpose classifier remotely, as it can be trained on a large set of short texts offline; moreover, the time to optimize the regression coefficients for the non-linear model is also practical for current mobile devices. Such an optimization is executed only sporadically by the users, typically when they feel that the estimation no longer reflects their own preferences.

The utility (loss) predictive model provides a building block that enables a tool to automatically select the level of location obfuscation to be applied to the check-ins. By predicting the utility loss for different obfuscation levels, such a tool can optimize the trade-off (i.e., find a sweet spot) between privacy and utility. As explained above, a straightforward solution is to optimize privacy (conversely utility) under a given threshold on utility (conversely privacy). Another solution is to optimize a consolidated criterion that takes into account both privacy and utility, as done by Olteanu et al. in their game-theoretical framework for strategic location-sharing [42].

6 Discussion and Limitations

The results presented in this work rely on a personalized user survey, conducted over Amazon Mechanical Turk, where participants were asked questions about their past check-ins on Foursquare, and on simple predictive models implemented in the WEKA toolkit. Although we tried to obtain unbiased responses from participants with a positive track-record and a minimum level of check-in diversity and used standard machine-learning techniques, our study still presents some limitations.

To start with, we did not perform any obfuscation on the user-generated text associated to a check-in. Such a text could contain information that can be used to identify the exact venue, even if other data is obfuscated on the semantic and geographic levels. Another limitation comes from the fact that our population sample included almost exclusively participants who are US residents, which could limit the applicability of our results to populations where information-sharing practices are significantly different. In addition, the results from our survey, and the features used in the predictive model, might be specific to Foursquare and not applicable to other LBSNs. The fact that we asked participants about their privacy preferences could introduce a bias towards a more privacy-conscious behavior. On the temporal dimension, we asked users to recall the purpose of check-ins that occurred as far as two years in the past (which makes it difficult for users to recall the context of their check-ins), thus allowing a judgment error on the users' part in case of bad recall due to recency and primacy effects [58]: Users tend to better recall situations that either happened recently or far in the past. This issue could be overcome by considering shorter periods of times (*e.g.*, one month in the past), or by including additional information

to help participants remember about the context of their check-in (*e.g.,* attached pictures). Moreover, the four coarse-grained purposes and the associated hierarchy, have not been extensively validated and could, in some cases, not reflect the actual purpose stated by the participants.

Regarding the performance of the models, which we believe to be good and promising, it could be further improved. For instance, by considering larger datasets (more users and more decisions per users; such data would be available should such a system be deployed), and by using more advanced models with fine-tuned parameters and more features. Because the size of the dataset is modest, the models might be biased; as for over-fitting, we limit the risks by using simple models with default parameters. Also, in order to avoid the problem of feature-selection (and the fact that the features are somewhat specific to Foursquare) we could use featureless techniques (more specifically, techniques that do not require the features to be manually provided), such as deep-learning techniques. But such techniques can only be used with very large training sets, making it unsuitable for our dataset and for our problem. It should be noted that the goal of our work is to show the feasibility of the approach rather than optimizing the performance of the models. Such optimizations are therefore left to future work.

Finally, quantifying utility is a difficult problem and our approach has some limitations. The fact that the notion of perceived utility is somewhat subjective could lead to different interpretations among the participants, despite the fact that we defined it formally as the "the extent to which the initial purpose of a check-in is still met" in the text of our survey. The use of a 5-point scale to quantify utility (with only the 1 and 5 options annotated) could also, to some extent, lead to different interpretations.

7 Dissemination of the Utility Dataset

In order to enable other research groups working on privacy protection in location check-in-based services to quantify utility, we release a sanitized version of our utility dataset, together with the associated predictive model (available at https://people.unil.ch/kevinhuguenin/datasets#utility). The sanitization process applied to the original dataset aims at protecting the privacy of the survey participants (in accordance with our commitment stated in the consent form approved by the ethics committee of EPFL). It includes removing uniquely identifiable survey participants (e.g., there was only one survey participant based in Canada), removing uniquely identifiable attributes (e.g., hometown of a user, text of a check-in), using broad categories for sensitive attributes (e.g., we used only three age ranges), *etc*. The sanitized utility dataset is available in the Attribute-Relation File Format (ARFF), an input file format used by the machine-learning tool WEKA that we employed for building our predictive models. It contains 13,376 instances (i.e., data points); each correspond to a survey response from a participant, for a given check-in and a given level of obfuscation (for each of her considered check-ins, a participant was presented with four obfuscated version and was asked to rate the utility on a scale from 1 to 5, as illustrated in Figure 2). In addition to the obfuscation level (Ls-Lg, Hs-Lg, Ls-Hg, Hs-Hg) and the utility rating (between 1 and 5), each instance contains fourteen features related to the user, the check-in, and the associated venue (see Table 6 for the complete description of the features). Despite the sanitization process, the released dataset still enables relatively accurate utility prediction: When using a J48 classifier trained and tested on the sanitized dataset following a 10-fold cross-validation approach (as shown in the WEKA-based Java program provided with the dataset), the mean prediction error is 0.89 (to be compared to an error of 0.52 for a classifier trained on the full original dataset and 1.9 for the baseline classifier).

8 Conclusion

In this paper, we study the users' motivations for checking in on a popular platform (Foursquare), and we design an automated mechanism to infer these motivations, in order to reduce the amount of unnecessarily disclosed details that are released by a check-in.

With our insights, we design and evaluate an efficient automated purpose inference mechanism. Furthermore, we re-use the output of the inference mechanism to build and evaluate a predictive model for utility, given the purpose of the check-in and the level of obfuscation. We show that a cost-sensitive classifier achieves a small mean prediction error, and we show that for more than 60% of users' check-ins, at least one of the proposed obfuscation methods can be used without significantly damaging their utility. This makes it possible for application and system developers, using generalization techniques, to incorporate privacy-preserving tools that have a negligible effect on the utility of the system. For instance, such a tool could choose the appropriate level of obfuscation (in terms of utility, based on– among other things –the inferred motivation behind the check-in) and either directly apply this level of obfuscation to the shared information or make a suggestion to the user and let her choose her preferred level of obfuscations.

Acknowledgments

We would like express our sincere gratitude to Nauman Shahid for his contribution to this project.

Sanitized Utility Dataset

The attributes contained in the sanitized utility dataset are listed in Table 6, together with short descriptions and the set of their possible values (large sets are truncated).

References

[1] 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 *NDSS*, 2015.

[2] K. Zickuhr, "Location-based services," Pew Research. 2013. http://www.pewinternet.org/files/old-media/Files/Reports/2013/PIP_Location-based%20services%202013.pdf. Last visited: Jan. 2014.

[3] S. Patil, G. Norcie, A. Kapadia, and A. J. Lee, "Reasons, rewards, regrets: Privacy considerations in location sharing as an interactive practice," in *SOUPS*, 2012, pp. 5:1–5:15.

[4] S. Patil, G. Norcie, A. Kapadia, and A. Lee, "Check out where i am!: location-sharing motivations, preferences, and practices," in *ACM CHI (Extended Abstracts)*, 2012, pp. 1997–2002.

[5] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman, "I'm the mayor of my house: Examining why people use foursquare – a social-driven location sharing application," in *ACM CHI*, 2011, pp. 2409–2418.

[6] S. Guha and J. Birnholtz, "Can you see me now?: location, visibility and the management of impressions on foursquare," in *MobileHCIs*, 2013, pp. 183–192.

[7] B. Ağır, K. Huguenin, U. Hengartner, and J.-P. Hubaux, "On the privacy implications of location semantics," *Proc. of the Privacy Enhancing Technologies (PoPETs)*, vol. 2016, no. 4, pp. 165–183, 2016.

[8] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *ACM MobiCom*, 2011, pp. 145–156.

[9] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *IEEE S&P*, 2011, pp. 247–262.

[10] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, "Identification via location-profiling in GSM networks," in *ACM WPES*, 2008, pp. 23–32.

[11] B. Lee, J. Oh, H. Yu, and J. Kim, "Protecting Location Privacy Using Location Semantics," in *ACM KDD*, 2011.

[12] O. Barak, G. Cohen, and E. Toch, "Anonymizing mobility data using semantic cloaking," *Pervasive and Mobile Computing*, vol. 28, no. C, pp. 102–112, Jun. 2016.

[13] 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.

[14] G. Wang, S. Y. Schoenebeck, H. Zheng, and B. Y. Zhao, ""will check-in for badges": Understanding bias and misbehavior on location-based social networks," in *ICWSM*, 2016, pp. 417–426.

[15] A. Acquisti and J. Grossklags, "Privacy and rationality in individual decision making," *IEEE Security Privacy*, vol. 3, no. 1, pp. 26–33, jan 2005.

[16] B. Knijnenburg and H. Jin, "The persuasive effect of privacy recommendations," in *SIGCHI Proceedings*, 2013, p. 16.

[17] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in GPS traces via uncertainty-aware path cloaking," in *ACM CCS*, 2007, pp. 161–171.

[18] K. Micinski, P. Phelps, and J. S. Foster, "An empirical study of location truncation on android," in *MoST*, 2013.

[19] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, "Protecting location privacy: Optimal strategy against localization attacks," in *ACM CCS*, 2012, pp. 617–627.

[20] H. Cramer, M. Rost, and L. E. Holmquist, "Performing a check-in: Emerging practices, norms and 'conflicts' in location-sharing using foursquare," in *MobileHCI*, 2011, pp. 57–66.

[21] H.-S. Kim, "What drives you to check in on facebook? motivations, privacy concerns, and mobile phone involvement for location-based information sharing," *Computers in Human Behavior*, vol. 54, pp. 397 – 406, 2016.

[22] S. S. Wang and M. A. Stefanone, "Showing off? human mobility and the interplay of traits, self-disclosure, and facebook check-ins," *Social Science Computer Review*, vol. 31, no. 4, pp. 437–457, 2013.

[23] Z. Zhu, U. Blanke, and G. Tröster, "Inferring travel purpose from crowd-augmented human mobility data," in *UrbIoT*, 2014.

[24] K. K. Rachuri, T. Hossmann, C. Mascolo, and S. Holden, "Beyond location check-ins: Exploring physical and soft sensing to augment social check-in apps," in *IEEE PerCom*, 2015, pp. 123–130.

[25] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *ACM MobiSys*, 2003, pp. 31–42.

[26] T. Jiang, H. J. Wang, and Y.-C. Hu, "Preserving location privacy in wireless LANs," in *ACM MobiSys*, 2007, pp. 246–257.

[27] R. Chow and P. Golle, "Faking contextual data for fun, profit, and privacy," in *ACM WPES*, 2009, pp. 105–108.

[28] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *ACM CCS*, 2013, pp. 901–914.

[29] A. Brush, J. Krumm, and J. Scott, "Exploring end user preferences for location obfuscation, location-based services, and the value of location," in *ACM UbiComp*, 2010, pp. 95–104.

[30] K. P. Tang, J. I. Hong, and D. P. Siewiorek, "Understanding how visual representations of location feeds affect end-user privacy concerns." in *ACM UbiComp*, 2011, pp. 207–216.

[31] B. Henne, C. Kater, M. Smith, and M. Brenner, "Selective cloaking: Need-to-know for location-based apps," in *PST*, 2013, pp. 19–26.

[32] J. Krumm, "Inference attacks on location tracks," in *Pervasive*, 2007, pp. 127–143.

[33] M. Decker, "Location privacy – an overview," in *IEEE ICMB*, 2008, pp. 221–230.

[34] M. Duckham, "Moving forward: location privacy and location awareness," in *ACM SPRINGL*, 2010, pp. 1–3.

[35] R. Shokri, "Privacy games: Optimal user-centric data obfuscation," *Proc. of the Privacy Enhancing Technologies (PoPETs)*, vol. 2015, no. 2, pp. 299–315, 2016.

[36] K. Fawaz, H. Feng, and K. G. Shin, "Anatomization and protection of mobile apps' location privacy threats," in *USENIX Security*, 2015, pp. 753–768.

[37] J. Yang, Z. Zhu, J. Seiter, and G. Tröster, "Informative yet unrevealing: Semantic obfuscation for location based services," in *ACM GeoPrivacy*, 2015, pp. 4:1–4:8.

[38] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in mechanical turk," in *ACM CHI*, 2010.

[39] W. Mason and S. Suri, "Conducting behavioral research on amazon's mechanical turk," *Behavior research methods*, vol. 44, no. 1, pp. 1–23, 2012.

[40] I. Bilogrevic, K. Huguenin, B. Ağır, M. Jadliwala, M. Gazaki, and J.-P. Hubaux, "A machine-learning based approach to privacy-aware information-sharing in mobile social networks," *Pervasive and Mobile Computing*, vol. 25, pp. 125–142, 2016.

[41] D. Meier, Y. A. Oswald, S. Schmid, and R. Wattenhofer, "On the Windfall of Friendship: Inoculation Strategies on Social Networks," in *ACM EC*, 2008, pp. 294–301.

[42] A.-M. Olteanu, K. Huguenin, M. Humbert, and J.-P. Hubaux, "The sharing game: Benefits and privacy implications of (co)-location sharing with interdependences," EPFL, research report, 2016. [Online]. Available: https://infoscience.epfl.ch/record/218755

[43] G. M. Sandstrom, N. Lathia, C. Mascolo, and P. J. Rentfrow, "Putting mood in context: Using smartphones to examine how people feel in different locations," *Journal of Research in Personality*, 2016, to appear.

[44] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1998.

[45] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing twitter" big data" for automatic emotion identification," in *PASSAT*, 2012, pp. 587–592.

[46] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment analysis," Stanford University, CS224N Project Report, 2009, http://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf.

[47] Sentiment 140, http://help.sentiment140.com/for-students, last visited: Mar. 2014.

[48] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *COLING (Posters)*, 2010, pp. 241–249.

[49] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification." in *Association for Computational Linguistics*, 2011, pp. 151–160.

[50] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREC*, 2010.

[51] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," in *Soviet physics doklady*, vol. 10, 1966, p. 707.

[52] K. Olejnik, I. Dacosta, J. S. Machado, K. Huguenin, M. E. Khan, and J.-P. Hubaux, "SmarPer: Context-aware and automatic runtime-permissions for mobile devices," in *S&P*, 2017, p. 19.

[53] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[55] N. Sadeh, J. Hong, L. Cranor, I. Fette, P. Kelley, M. Prabaker, and J. Rao, "Understanding and capturing people's privacy policies in a mobile social networking application," *Personal Ubiquitous Computing*, vol. 13, no. 6, pp. 401–412, Aug. 2009.

[56] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov, "Android permissions remystified: A field study on contextual integrity," in *USENIX Security*, 2015, pp. 499–514.

[57] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[58] J. Murdock and B. Bennet, "The serial position effect of free recall," *Journal of experimental psychology*, vol. 64, no. 5, p. 482, 1962.

Kévin Huguenin is an assistant professor at HEC Lausanne–UNIL, which he joined in 2016. Prior to that, he held a permanent researcher position at LAAS-CNRS, France. He also worked as a post-doctoral researcher at EPFL and at McGill University. He earned his Ph.D. in computer science from the Université de Rennes and Inria, France, in 2010 and his M.Sc. degree from École Normale Supérieure de Cachan and the Université de Nice, France, in 2007. His research interests include information security and privacy.

Igor Bilogrevic is a research scientist at Google, which he joined in 2014. He earned his Ph.D. on the privacy of context-aware mobile networks from EPFL in 2014. From 2010 until 2012, he worked in collaboration with the Nokia Research Center on privacy in pervasive mobile networks, encompassing social community and location privacy and information-sharing. In 2013, he spent the summer at PARC (a Xerox Company) working on topics related to private data analytics. His main research interests lie at the frontiers between privacy, security and user experience.

Joana Soares Machado recently earned her M.Sc. degree in communication systems, with a specialization in networking and mobility, from EPFL. While working towards the M.Sc. degree, she did a 6-month internship at CERN. She earned her B.Sc. degree in telecommunications and computer engineering from ISCTE - University Institute of Lisbon, Portugal, in 2013. She is the vice-president of the EPFL IEEE student branch. Her research interests include privacy and machine learning.

Stefan Mihaila is a software engineer currently employed in London, UK. He earned his M.Sc. degree in computer science from EPFL, Lausanne, Switzerland, and his B.Sc. degree in computer science from University Alexandru Ioan Cuza, Iași, Romania. While working towards the M.Sc. degree, he conducted projects on various topics including privacy, cryptography, and databases in different labs at EPFL and he did a 6-month internship at Twitter. His research interests include machine learning, distributed systems and cryptography.

Reza Shokri is an assistant professor in the computer science department at NUS, Singapore. Prior to that, he was a post-doctoral researcher at Cornell Tech, NY, USA, UT Austin, TX, USA, and at ETH Zurich, Switzerland. He earned his Ph.D. on the quantification and protection of location privacy from EPFL in 2013 and his M.Sc. in software computer engineering from the University of Tehran, Iran, in 2007. His research focuses on computational privacy.

Italo Dacosta is a post-doctoral researcher at EPFL, which he joined in 2014. Prior to that, he worked as a post-doctoral researcher at KU Leuven, Belgium. He earned his Ph.D. in computer science and his M.Sc. degree in information security from the Georgia Institute of Technology, USA. He earned his B.Sc. degree in electronic and communication engineering from the Universidad de Panama, Panama, in 2002. He is also a former Fulbright grant recipient. His research interests include privacy enhancing technologies (PETs), identity management and network and mobile security.

Jean-Pierre Hubaux is a full professor at EPFL, which he joined in 1990. His current research activity is focused on privacy, notably in pervasive communication systems and online social networks. He has recently started research activity in genomic privacy, in close collaboration with geneticists. In 2008, he completed a graduate textbook, entitled Security and Cooperation in Wireless Networks, with Levente Buttyan. He held visiting positions at the IBM T.J. Watson Research Center and at UC Berkeley. He is a fellow of both the ACM and IEEE.

Table 6
Description of the attributes present in the sanitized utility dataset.

name	**description**	**values**
`user_age`	age of the user	{"[18-26)", "[26-36)", "[36-)"}
`user_gender`	gender of the user	{"male", "female"}
`checkin_part_of_day`	time of check-in time	{"morning", "afternoon", "evening", ...}
`checkin_day_of_week`	day of check-in	{"Monday", "Tuesday", ...}
`checkin_has_text`	text attached to the check-in	{"yes", "no"}
`checkin_has_picture`	pictures attached to the check-in	{"yes", "no"}
`checkin_has_colocation`	other users tagged in the check-in	{"yes", "no"}
`venue_nb_checkins`	number of check-ins at the venue	{"[0-100)", "[100-500)", "[500-1000)", ...}
`venue_distance_home`	distance between the venue and the user's home [km]	{"[0-100)", "[100-1000)", ...}
`venue_same_city`	venue is in the same city as the user's home	{"yes", "no"}
`venue_same_state`	venue is in the same state as the user's home	{"yes", "no"}
`venue_same_country`	venue is in the same country as the user's home	{"yes", "no"}
`venue_type_root1`	first ancestor of the venue type (starting from the root[5])	{"Food", "Arts & Entertainment", ...}
`venue_type_root2`	second ancestor of the venue type (starting from the root[5])	{"Burger Joint", "Movie Theater", ...}
`obfuscation`	level of obfuscation	{"Ls-Lg", "Ls-Hg", "Hs-Lg", "Hs-Hg"}
`utility`	utility rating	{1,2,3,4,5}