# Predicting Users' Motivations behind Location Check-Ins and Utility Implications of Privacy Protection Mechanisms

Igor Bilogrevic, Kévin Huguenin, Stefan Mihaila, Reza Shokri, Jean-Pierre Hubaux

# Predicting Users' Motivations behind Location Check-Ins and Utility Implications of Privacy Protection Mechanisms


Igor Bilogrevic* Google Switzerland ibilogrevic@google.com

Kévin Huguenin* LAAS-CNRS France huguenin@laas.fr

Stefan Mihaila EPFL Switzerland stefan.mihaila@epfl.ch

Reza Shokri† University of Texas, Austin USA shokri@cs.utexas.edu

Jean-Pierre Hubaux EPFL Switzerland jean-pierre.hubaux@epfl.ch



***Abstract*—Location check-ins contain both geographical and semantic information about the visited venues, in the form of tags (e.g., "restaurant"). Such data might reveal some personal information about users beyond what they actually want to disclose, hence their privacy is threatened. In this paper, we study users' motivations behind location check-ins, and we quantify the effect of a privacy-preserving technique (i.e., generalization) on the perceived utility of check-ins. By means of a targeted user-study on Foursquare (N = 77), we show that the motivation behind Foursquare check-ins is a mediator of the loss of utility caused by generalization. Using these findings, we propose a machine-learning method for determining the motivation behind each check-in, and we design a motivation-based predictive model for utility. Our results show that the model accurately predicts the loss of utility caused by semantic and geographical generalization; this model enables the design of utility-aware, privacy-enhancing mechanisms in location-based social networks.**


## I. Introduction

Online social networks (OSNs), such as Facebook and Foursquare, allow their users to share location information with each other. Such a feature is quite popular, as 30% of users attach locations to their posts [36]. The reason for sharing locations include the desire to connect with users' social circles and to project an interesting image of themselves [27], [28], thus achieving a goal greater than simply disclosing geographical information [12], [21].

By checking-in to a place or an event, on so-called location-based social networks (LBSNs), 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 to 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" (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 might lead to the exposure of additional private information beyond what they intended to share. A collection of location check-ins by a set of users can lead to their re-identification and also an inference of more personal information (e.g., complete location trace, co-travelers, activities) [7], [30], [35]. The risks are even higher when users share semantic information as well. For example, activity patterns can be learned at the semantic level (e.g., users go to cinemas after dining in restaurants) and subsequently used to better track users' locations.

To protect their privacy, users can obfuscate their location information, both at the geographical and semantic levels. For example, a user can generalize[1] the semantic information of the venue by sharing, for example, "restaurant" instead of "burger joint". The user can also generalize the geographical location of the venue by sharing, for instance, the city instead of the full address of the venue. Location obfuscation decreases the chances that a curious entity can track the location and activities of the user over time, hence it increases the user's privacy. However, this might come at the cost of a reduction in her perceived quality of service (i.e., utility).

Because it is difficult for users to estimate the privacy risks that stem from location sharing (it usually requires to perform statistical inference [30]) and because it would be cumbersome for users to manually select the level of obfuscation to apply to each of their check-ins, *automatic* obfuscation mechanisms are needed (note that automatically generated privacy recommendations are valuable as well [18]). In order 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., [30]. However, few studies address the utility loss due to location obfuscation for particular location-based services [15], [23], or the utility loss in a formal framework for finding the optimal balance between utility and privacy [31]. Despite these studies, there is no methodology for modeling and predicting the perceived utility loss that stems from the use of

*This work was carried out while the author was with EPFL.

†This work was carried out while the author was with ETH Zurich.




NDSS '15, 8-11 February 2015, San Diego, CA, USA





http://dx.doi.org/10.14722/ndss.2015.23032


[1] In this paper, 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 left to future work.

obfuscation mechanisms in location-based social networks for each individual check-in (for each individual user). This paper provides such a methodology in order to design automatic personalized location privacy protection mechanisms.

The problem of predicting a user's perceived utility loss due to obfuscation is highly 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.[2] This determines 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.

In order to find the right balance between the level of obfuscation and the utility requirements of each user, we use machine learning algorithms that, given some features about a check-in (and the user's behavior), predict her motivation for this check-in and her perceived utility loss for each level of (geographical and semantic) location obfuscation. The result of our algorithm is a personalized utility loss function. We implement and test our methodology on the results of an online survey involving 77 Foursquare users (with 45 check-ins per user). We can predict the purpose of the check-ins (among 13 pre-selected purposes) with a raw correct classification rate of 43% and the effect of obfuscation on utility (on a scale from 1 to 5) with a mean prediction error of 0.66.

The results of our survey also shed light on the effects of location obfuscation mechanisms on the perceived utility 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 impact on the perceived utility, compared to geographic obfuscation (e.g., reporting the city instead of the full address).

In summary, our contributions are as follows:

1) We present the first methodology, to the best of our knowledge, 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 loss function that can be used as a building block for designing usable location privacy-protection mechanisms. Such mechanisms could automatically choose the *best* 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.

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

The rest of the paper is organized as follows. After discussing the related work in Section II, we present the methodology of our study in Section III, which includes an online survey with Foursquare users, and the definition of the motivation and utility inference framework. Subsequently, we present quantitative results, by discussing both descriptive statistics and performance values of our motivation classifier and utility model in Sections IV and V respectively. We then discuss the limitations of our study, conclude the paper and give directions for future work in Section VI.

## II. 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.

### A. Motivations behind Location Sharing

Recently, several works investigated the users' motivations for disclosing their locations in online social networks. Patil *et al.* [27], [28] carried out two online user-studies, with 401 and 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 to appear "cool", rather than to pointing to a specific geographical location. Similarly, results presented in [12], [21] 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 [27], [28] as the default options available to users for selecting the main purposes of their check-ins. In order not to restrict users to one of the predefined choices, we also offer them the option for entering a purpose that is not present in the predefined list. Cramer *et al.* [5] performed an in-deep 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 that 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. 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.

### B. 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 [1], [4], [11], [15], [17].

Brush *et al.* [3] studied the users' preferences and concerns for several such algorithms by visually 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.* [32] presents the users with three different visualizations of their past shared locations and studies their effect on the end-user privacy. They show that, based on the type of visualization, the users expressed diverging attitudes towards 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 [14], [23]. In particular, Micinski *et al.* [23] 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 less precise locations does not substantially hinder their functionality. A more encompassing approach, taken by Henne *et al.* [14], 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 them caused by the information they share, there are several attempts to formalize the desirable location privacy requirements that obfuscation mechanisms should fulfill and the metrics to quantify them. Examples of such pieces of work are Krumm [19], Decker [8], and Duckham [9]. In a follow-up of these works, Shokri *et al.* provide a framework [30] to quantify location privacy, and a game-theoretic methodology [31] 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. Few studies that include utility aspects of location obfuscation mechanisms only reflect the application dimension of it, for example, by measuring the fraction of restaurants that a user misses, or the error of traffic information due to location perturbation [15], [23]. Our work completes this line of studies, by providing a methodology to design user-centric utility functions for location check-ins.

## III. 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 into 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 to 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.[3] Each venue is associated with a street address and a semantic tag (from a predefined set of tags, organized as a tree). 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 their purpose 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.

### A. Participants and Remuneration

To recruit participants, we made use of the Amazon Mechanical Turk (MTurk) platform, which allowed us to draw candidates from a pool of users with diverse backgrounds 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 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, (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 to 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 only considered 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 non-serious workers from participating in our survey.

Our survey is based on the participants' actual check-ins on Foursquare posted over the last 24 months (that 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 grant us access to their Foursquare data, we rewarded them with a fixed amount of money (US $4.5 per HIT [2], [22]). At the end of the study, the average per-hour remuneration for the participants was US $8.50. The total budget for the experiment was $600.

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

## B. 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[4], 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 9 questions for each of the 45 check-ins, 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 1).[5]

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 [27], [28] and our internal experiment) or to specify a different one in free-text.

We considered two levels of obfuscation (low and high), both at the geographical and at the semantic levels. Geographic obfuscation reveals only some of the geographic information (among 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 ancestors). The four combinations of obfuscation levels are defined as follows and are illustrated on a sample venue in Table I:

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.

[4]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.

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

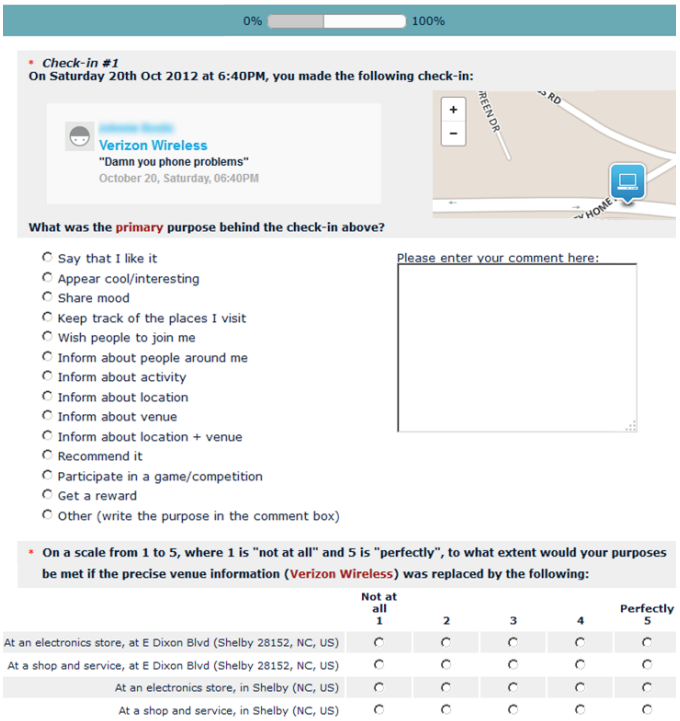


Fig. 1. Screenshot of our online survey website. Participants are presented with some of their own past Foursquare check-ins and they are asked some questions about the purpose of their check-ins and the effect of (geographical and semantic) location obfuscation on their perceived utility. For privacy reasons, we blurred the name of the participant.

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 I 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. Figure 1 shows a 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.

TABLE I. 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) |

### C. 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 is 29±6 years, where the oldest and youngest participants were of age 50 and 19, respectively. 43% were male, and participants were almost exclusively based in the US (96%). The other participants came from Canada (1), Norway (1) and Israel (1). Regarding their occupation, 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").

### D. 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 99% of all 3465 check-ins. Figure 2 shows the distribution over the participants' purposes for their check-ins. We can see that, among the top four 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) or to personal record-keeping purposes, which corroborates the results obtained by [27], [28]. The purpose of informing about the actual location was only selected for less than 9% of the check-ins.

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.

### E. 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

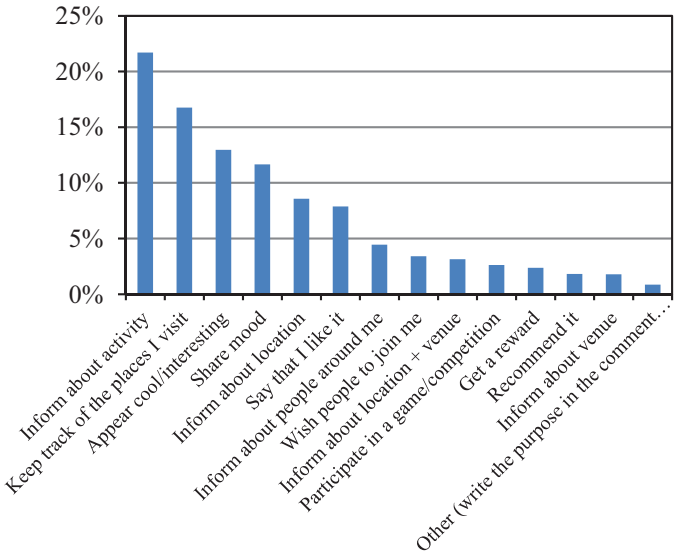


Fig. 2. Proportion of purposes for the users' check-ins. The top four purposes, which account for 63% of the total, represent only high-level social and personal goals. Informing about the actual location is only the 5th most frequent purpose, selected in less than 9% of the cases.

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 I) 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.

*1) Utility vs. Obfuscation (in General):* In order to study the general relationship between utility and obfuscation, we group the check-ins according to 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 3.[6]

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-

[6]The differences among the averages of the four obfuscation levels are statistically significant, both pairwise and globally ($\chi^2$ test of homogeneity, $p < .01$).

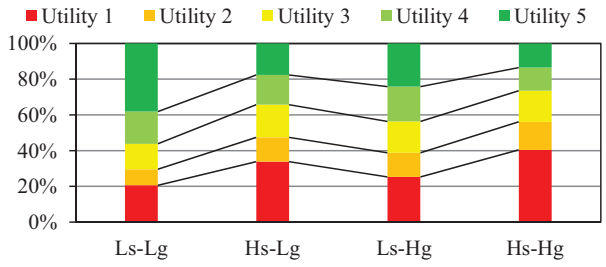


Fig. 3. Proportion of check-ins with their perceived utility, for different levels of geographical and semantic obfuscation. A utility of 1 means that the purpose of the check-in is not met at all after the obfuscation, whereas a utility of 5 means that the purpose is met perfectly after the obfuscation. Perceived utility decreases with the level of obfuscation; semantic obfuscation has a stronger (negative) effect on utility.

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 clear 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.

*2) Utility vs. Obfuscation (given the Purpose):* Figure 4 shows the participants' utility scores for check-ins, grouped according to their purpose: "Inform about activity" (Figure 4a), "appear cool/interesting" (Figure 4b), and "wish people to join" (Figure 4c).

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 the geographical one. In particular, compared to the Ls-Lg scenario, the lowest utility score increases from 19% to 40% (+111%), when increasing the semantic obfuscation; 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.[7]

For check-ins with the purpose of appearing cool/interesting (Figure 4b), the utility scores exhibit lower variations as compared to Figure 4a and more in accordance with the general motivation-utility results shown in Figure 3.[8] An interesting result is shown by Figure 4c, 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 5): 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 the (inferred) purposes of check-ins to predict the utility loss caused by the use of different obfuscation techniques.

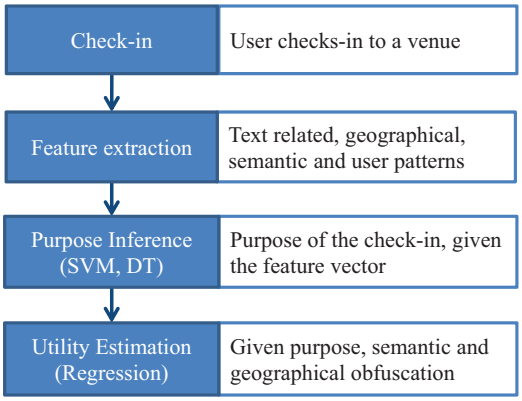


Fig. 5. Workflow of the the utility model framework, including the purpose inference stage.

## IV. 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. Moreover, to capture the dependencies between the structured and unstructured features, we also create several hybrid features.

### A. From Check-ins to Features

The three types of the aforementioned features are combined in a single feature vector that will be fed to the machine learning algorithm, in order to derive the most likely purpose for each check-in. Hereafter we describe all the different components of the feature vector.

[7] $p < .01$

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

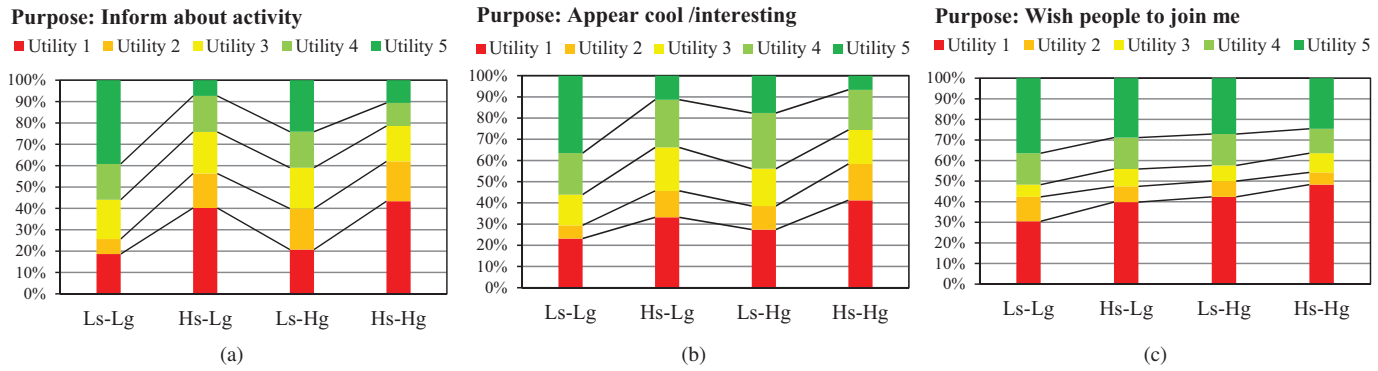


Fig. 4. Proportion of check-ins with their perceived utility, for different levels of geographical and semantic obfuscation, according to their purpose.

*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 per venue, and complete address. Moreover, we extract a venue's ancestors in the semantic hierarchy, as well as the user's age, number of total check-ins, occupation and gender.

*2) Unstructured Text Features:* Based on 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) [33] and the sentiment (such as positive or negative) [10]. 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[9], and we used a Naïve Bayes classifier (trained on relevant short-texts [29], [33]) 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 [6], [16], [25]. We also include some features that capture the presence of specific keywords 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).

*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 [20] 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.

### B. 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. Figure 5 shows the work flow of the entire inference process. We experimented with different state-of-the-art classifiers (including a Support Vector Machine (SVM) classifier with a Gaussian kernel function, trained with the Sequential Minimal Optimization (SMO) algorithm, a Random Forests classifier using up to 100 trees of up to 10 features and a Logistic classifier with the LogitBoost algorithm). Our results are obtained using the well-established WEKA toolkit [13], based on 10-fold cross validation. We use the data obtained through our survey as ground-truth to train the classifier and validate the results.

Table II shows the performance of our purpose inference classifier (using Random Forests, which give the best results on our dataset) in the form of a confusion matrix. These results are obtained on 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. As a global performance metric, we use the Correct Classification Rate (CCR), that is 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%; 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, 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 (+95%). This number is relatively high considering the high number of possible purposes. Note that misclassifications have different levels of severity (classifying a check-in with purpose "Recommend it" as "Say I like it" can be considered better, in terms of similarity, than classifying it as "Get a reward"). Hence, 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 55% (58% among the check-ins for which a secondary purpose was reported). As part of future work, we intend to investigate further the severity of misclassifications. In particular, we will consider hierarchical models for the different purposes (e.g., "Inform about venue" and "Inform about activity" could be clustered in the "Inform" meta-purpose).

We also look at the precision and recall for each class (i.e., purpose). The precision for purpose (a) is defined as

[9]Available from http://www.nltk.org/.

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 classification.

TABLE II. CONFUSION MATRIX FOR THE 13-CLASS PURPOSE CLASSIFIER, WITH THE PER-CLASS 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. BY USING THE CHECK-IN INFORMATION, THE CLASSIFIER CAN INFER THE CORRECT PURPOSE IN 43% OF THE CASE, WHICH IS ALMOST TWICE AS GOOD AS THE BASELINE (+95%).

| ↓ Classified as → | | (a) | (b) | (c) | (d) | (e) | (f) | (g) | (h) | (i) | (j) | (k) | (l) | (m) | Total (%) | Prec. | Rec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Inform about location | (a) | 93 | 2 | 90 | 10 | 9 | 19 | 4 | 10 | 10 | 34 | 2 | 3 | 11 | 297 (9%) | 42% | 31% |
| Recommend it | (b) | 1 | 4 | 16 | 6 | 3 | 14 | 0 | 4 | 2 | 7 | 5 | 0 | 1 | 63 (2%) | 14% | 6% |
| Inform about activity | (c) | 40 | 4 | 451 | 51 | 14 | 57 | 5 | 26 | 11 | 80 | 3 | 5 | 5 | 752 (22%) | 46% | 60% |
| Appear cool/interesting | (d) | 7 | 1 | 79 | 177 | 16 | 45 | 0 | 27 | 9 | 74 | 5 | 3 | 6 | 449 (13%) | 40% | 39% |
| Inform about people around | (e) | 6 | 2 | 40 | 21 | 30 | 19 | 2 | 6 | 5 | 14 | 1 | 3 | 5 | 154 (4%) | 26% | 19% |
| Share mood | (f) | 14 | 3 | 85 | 54 | 12 | 159 | 1 | 22 | 7 | 39 | 2 | 1 | 5 | 404 (12%) | 38% | 39% |
| Inform about venue | (g) | 6 | 0 | 20 | 1 | 2 | 8 | 4 | 4 | 3 | 12 | 0 | 1 | 1 | 62 (2%) | 16% | 6% |
| Say that I like it | (h) | 10 | 2 | 64 | 42 | 8 | 26 | 2 | 67 | 8 | 35 | 4 | 3 | 2 | 273 (8%) | 34% | 25% |
| Wish people to join me | (i) | 10 | 2 | 17 | 11 | 6 | 10 | 2 | 5 | 41 | 10 | 1 | 2 | 1 | 118 (3%) | 38% | 35% |
| Keep track of the places I visit | (j) | 19 | 4 | 77 | 44 | 10 | 47 | 5 | 23 | 9 | 324 | 8 | 7 | 4 | 581 (17%) | 49% | 56% |
| Get a reward | (k) | 3 | 3 | 5 | 7 | 0 | 4 | 0 | 2 | 0 | 11 | 41 | 6 | 0 | 82 (2%) | 53% | 50% |
| Participate in a game | (l) | 4 | 0 | 6 | 9 | 1 | 4 | 0 | 1 | 0 | 7 | 6 | 53 | 0 | 91 (3%) | 61% | 58% |
| Inform about location + venue | (m) | 9 | 1 | 27 | 11 | 3 | 7 | 0 | 2 | 2 | 14 | 0 | 0 | 33 | 109 (3%) | 45% | 30% |

| | |
|---|---|
| **Correct classification rate** | **43%** |
| **Correct classification rate (baseline)** | **22%** |

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 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, and thus 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 likely because this purpose is quite generic, and because the proportion of such check-ins is too low to efficiently learn meaningful patterns while training.

Finally, we consider the sorted lists of purposes returned by the classifier (instead of looking at just the first purpose returned) and we look at the position (or rank) of the actual purpose of the check-ins in this list. Figure 6 shows the histogram and the cumulative distribution function of the rank. It can be observed that, in 60% 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. 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).

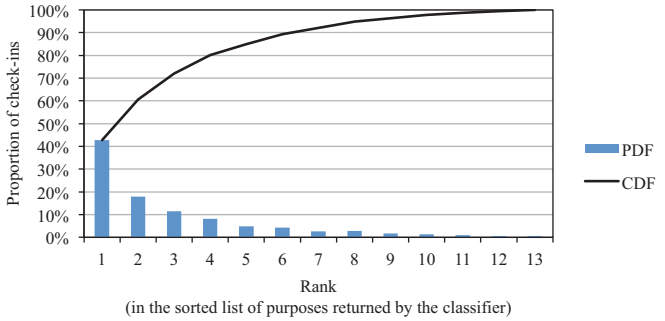


Fig. 6. Rank of the actual purposes of the check-ins in the sorted list of purposes returned by the classifier.

## V. PURPOSE-BASED 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.

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 $u = f(m, \mathbf{o}, \mathbf{k}) \in [1, 5]$ of a check-in, given the purpose $m \in \{1, \ldots, 13\}$ of the check-in, the level of (semantic and geographical) obfuscation $\mathbf{o} = (o_s, o_g)$, where $o_s, o_g \in \{1, 2\}$ are the semantic and geographical obfuscation levels, and where $\mathbf{k} = (k_1, \ldots, k_n)$ are characteristics of the venue and of the user, such as her age group, the probability of visiting a given category of a venue and general privacy concerns. $o_s = 1$ and $o_s = 2$ mean that a *low*, respectively *high*, level of semantic obfuscation has been applied. We have a total of four different obfuscation scenarios for each check-in.

We split the analysis of the utility model into two parts. First, we study the relationship between utility, obfuscation and motivation in a model that uses the actual ground-truth data about the purposes of the users; in this model, we do not consider the output of the machine-learning classifier for the purpose of the check-ins. This way, we can study directly the relationship between utility and purpose. Second, we study this relationship but, instead of using the actual purposes of the check-ins, we rely on the output of the automated classifier. This enables us to compare the two models, where the difference is that, in the first case, users provided the actual purpose of each check-in, whereas in the second case, the purpose was inferred automatically, without asking the user for her input. We compute the regression coefficients and the related statistics in the R software, as well as by using WEKA.

With respect to the regression functions, we compare a linear regression function against a non-linear one (M5P model tree technique [34], by using the WEKA toolkit). First, we define a linear model that takes into account the purpose of the check-ins, the semantic and geographic obfuscation levels, and characteristics of the users. To construct function, we first

create a set of 13 dummy binary variables $\{m_d\}_{d=1}^{13}$ to encode the 13 possible purposes; similarly, we generate 4 dummy variables for the time of the day to encode 5 different possibilities (morning, noon, etc.). Moreover, we take into account the correlation and mutual dependence between obfuscation and purpose by having them appear as factors in the regression function. In the end, we obtain 13 binary variables for the purpose ($m_d$), 3 variables for the obfuscation levels ($o_s \cdot o_g$, $o_s$, $o_g$) and 23 for the venue and user characteristics, where each of the 7 variables $k_i$ is converted to several binary variables $k^{(j)}$. The linear regression function is defined as:

$$u_{\text{lin}}(m, \mathbf{o}, \mathbf{k}) = a_0 + (\sum_{d=1}^{13} b_d \cdot m_d) + (\sum_{j=1}^{23} c_j \cdot k^{(j)}) + e_0 \cdot o_s \cdot o_g + e_1 \cdot o_g + e_2 \cdot o_s$$

where $a_0, b_i, c_i, e_i$ are the coefficients that we estimate by using the least squares method. Second, we use the WEKA toolkit in order to evaluate the non-linear model and ascertain whether there is a significant difference between the two models. We expect the non-linear model to perform better than the linear one; however, the linear model will provide us with results that can be interpreted on a per-feature basis, and will allow us to compare their relative coefficients in the regression function, as shown hereafter.

### A. Linear Model of Utility vs. Purpose

*1) Actual Purpose vs. Utility:* In this scenario, we consider the actual reported purposes of the check-ins when optimizing the regression coefficients. Hence, the purpose vector $\mathbf{m}$ is a binary vector, where there is at most one occurrence of the value 1 for each such vector.

The linear model achieves a $R^2 = 0.20$, with a mean error of 1.19 over the range $[1, 5]$, and a $p$-value $< .01$. In terms of motivation coefficients, we observe that the largest has a value $-0.63$ ($p < .01$) for the purpose "inform about people around me", whereas the only one that has a positive effect on utility is "say that I like it", with a value of $0.41$ ($p < .01$). In general, most motivation predictors are significant, although they have a relatively small contribution ($< -0.3$) on the overall utility. With respect to the coefficients for the semantic and geographic obfuscation, we observe that both of them have a negative effect on utility ($-0.73$ and $-0.40$, respectively). However, there is also a clear difference in their magnitudes: The one for the semantic obfuscation is almost two times higher than the one for the geographic obfuscation. In this respect, our findings corroborate the prior results in [27], [28], by quantifying the impact on the utility of both different motivations and levels of detail [26] for real Foursquare check-ins.

Overall, the regression results show that when the actual purposes are known, the linear model does not achieve good results in terms of fit, and it still maintains a modest mean error over the considered range. It shows, however, how some of the motivations and obfuscation parameters are indeed significant for the prediction of utility.

*2) Inferred Purpose vs. Utility:* In this scenario, the actual purpose of the check-in is not known. As a consequence, the purpose vector is not a binary vector anymore but it contains probabilities, as they are output by the SVM purpose classifier of the previous stage. On the one hand, this provides less certainty about the actual purpose of the check-in; on the other hand, it enables a linear combination of purposes to be expressed in the regression function, instead of a single one.

The regression results for this scenario show that, overall, the linear model achieves a slightly better fit ($R^2 = 0.21$) and a slightly lower mean error (1.18), where $p < .01$. In terms of coefficients of the purpose parameters, we observe that they are all positive and larger than 8, as the purpose vector contains the probability distribution over purposes, and thus larger coefficients can be used for the regression. The largest predictor is the same as in the previous case, i.e., "inform about people around" (value of 32.45, $p < .01$). Moreover, the coefficients of the other parameters (obfuscation and user features) are similar to the previous case as well. The intercept is negative at $-10.8$.

Compared to the case where the actual purposes are known, the inferred purposes achieve overall similar results, although we observe a slight improvement of 5% in terms of overall fit of the model for the case where the purpose classifier is used. This suggests that better results can be achieved for the linear model, by allowing for a larger flexibility in terms of purposes. We believe that, in practice, this is to be expected as users who check into places usually do so for a combination of purposes, rather than a single one. In our dataset, we also collected information about an optional secondary purpose, but we obtained too few entries for such an information.

### B. Non-linear Model of Utility vs. Purpose

In order to overcome possible limitations of the linear model, we compared the previous results with those obtained by using a non-linear model based on the model tree technique M5P [34]. This model produces a tree of regression models, where linear regression functions are found at the nodes of the tree. We performed the regression over all the check-ins in WEKA, using 10-fold cross validation.

We first consider the case where the actual purposes are known. The regression produces significantly better results in terms of mean absolute error of prediction, which is 0.66 compared to 1.19 of the linear model (-56%), by taking into account 362 rules present in the tree. As expected, the non-linear model performs better than the linear one, as the M5P model is better able to model the complex subtleties of the users perceived utility. The correlation coefficient of the overall model is also relatively high (81%). In particular, we observe that the users' age is the first attribute that is considered in the M5P output tree, i.e., the age provides the largest reduction in the error of the utility regression function: For participants who are less than 33 years old, the subsequent attribute is the level of semantic obfuscation; however, for participants that are older, the subsequent attribute is the frequency of visiting the second-ancestor of the check-in venue. This finding shows how participants that belong to distinct age groups seem to use different criteria when evaluating the utility of check-ins after they are obfuscated. As part of future work, we intend to further study the relationship between motivation-based features and demographic ones, by means of a semi-structured interviews in addition to online surveys. For the case where the purposes are inferred and not known, we observe a

slightly higher mean absolute error (0.68) and a comparable correlation coefficient (80%), for a slightly higher number of rules (442) of the tree.

Overall, the regression results suggest that the non-linear model should be preferred to the linear one in terms of errors in the prediction, as well as for the fit of the function. In our utility function, we asked participants to rate the utility of the obfuscated check-ins on a discrete 5-point scale; however, by allowing a continuous interval for utility, the non-linear model should perform even better. This would enable automated purpose inference mechanisms to work together with utility estimation techniques, in order to select the level of semantic and geographic obfuscation that yields the best utility for any given check-in. In the following subsection we provide one way to implement this.

### C. Privacy/Utility Trade-Off

In our study, one straightforward way to take the privacy of check-ins into account is to associate it with the different obfuscation level. In particular, we 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).

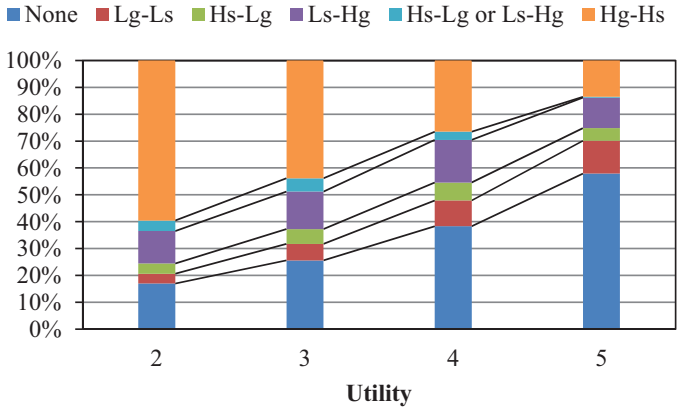


Fig. 7. Proportion of check-ins that can be obfuscated to the highest level among the four semantic and geographical combinations, for a given utility value in the interval $\{2, \ldots, 5\}$. If no combination of obfuscation meets the utility value, the highest obfuscation combination is set to "None", i.e., we keep the full details of the check-in.

For each utility value in $\{2, \ldots, 5\}$, Figure 7 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 (and thus 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 in 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 allow 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 locally, 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 (less than 9 seconds for the full set of $3465 \cdot 4$ check-in variants in our dataset). Such an optimization is executed only sporadically by the users, typically when they feel that the estimation no longer reflects their own preferences.

### D. Limitations

The results presented in this work rely on a personalized user-study, conducted over Amazon MTurk, where participants were asked questions about their past check-ins on Foursquare. Although we tried to obtain responses from participants with a positive track-record and a minimum level of check-in diversity, our study still presents some limitations.

Notably, 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 may be specific to Foursquare (and not applicable to other LBSNs). On the temporal dimension, we asked users to recall the purpose of check-ins that occurred as far as 2 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 [24]: 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). Finally, the use of a 5-point scale to quantify utility (with only the 1 and 5 options annotated) could lead to different interpretation between participants.

We intend to overcome some of the aforementioned limitations by integrating a larger number of participants through more diverse advertisement campaigns that, in addition to MTurk, include a broader set of people from other countries.

## VI. 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 and exploit these motivations, in order to reduce the amount of excessive details that are released by a check-in. First, we show that the purposes of check-ins play a significant role in determining their utility for the users, after we remove or replace some details on the semantic and geographic levels. In particular, we show that obfuscating (or reducing) information on the semantic level has a significantly more negative effect on the utility of the check-ins, compared to obfuscating on the geographic level.

By exploiting these insights, we design and evaluate an automated purpose inference mechanism, showing that it achieves a performance that is two times better than the baseline. Furthermore, we re-use the output of the inference mechanism to build and evaluate a regression model for utility, given the purpose of the check-in and the level of obfuscation. We show that a non-linear characterization of utility achieves a small prediction error (0.68 over the range $[1, 5]$), 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 usability of the system, yet provide a higher level of privacy to the users. 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 the level of obfuscation she prefers.

Beyond helping model the perceived utility, inferring the purposes of individual location check-ins can reveal useful to create new features on LBSNs. For example, users could be offered the "directions to the venue" feature for check-ins which purpose is "Wish people to join me" or offered to share a group picture for check-ins which purpose is "Inform about people around me". More generally, the classification of the check-ins (wrt their purposes) could be used to automatically adjust the way the check-in history is presented to the users.

As part of future work, in addition to overcoming some of the limitations we discussed, we plan to provide further insight into behavioral patterns and provide explanations for the regression models, by collaborating with experts from social psychology at partner institutions. Moreover, we intend to study the differences, in terms of check-in behaviors (and the implications on the perceived utility of check-ins), between different LBSNs. Finally, we plan to run a trial (based on an mobile app that allows users to obfuscate their check-ins) in order to assess the potential of our approach in the wild.

## Acknowledgments

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

## References

[1] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *CCS'13: Proc. of the 20th ACM Conf. on Computer and Communications Security*, 2013, pp. 901–914.

[2] I. Bilogrevic, K. Huguenin, B. Agir, M. Jadliwala, and J.-P. Hubaux, "Adaptive information-sharing for privacy-aware mobile social networks," in *UbiComp '13: Proc. of the 2013 ACM Int'l joint Conf. on Pervasive and Ubiquitous Computing*, 2013, pp. 657–666.

[3] A. Brush, J. Krumm, and J. Scott, "Exploring end user preferences for location obfuscation, location-based services, and the value of location," in *UbiComp'10: Proc. of the 12th ACM Int'l Conf. on Ubiquitous Computing*, 2010, pp. 95–104.

[4] R. Chow and P. Golle, "Faking contextual data for fun, profit, and privacy," in *WPES'09: Proc. of the 8th ACM Workshop on Privacy in the Electronic Society*, 2009, pp. 105–108.

[5] H. Cramer, M. Rost, and L. E. Holmquist, "Performing a check-in: Emerging practices, norms and 'conflicts' in location-sharing using foursquare," in *MobileHCI'11: Proc. of the 13th Int'l Conf. on Human Computer Interaction with Mobile Devices and Services*, 2011, pp. 57–66.

[6] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *COLING'10: Proc. of the 23rd Int'l Conf. on Computational Linguistics: Posters*, 2010, pp. 241–249.

[7] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, "Identification via location-profiling in GSM networks," in *WPES'08: Proc. of the 7th ACM Workshop on Privacy in the Electronic Society*, 2008, pp. 23–32.

[8] M. Decker, "Location privacy – an overview," in *ICMB'08: Proc. of the 2008 7th IEEE Int'l Conf. on Mobile Business*, 2008, pp. 221–230.

[9] M. Duckham, "Moving forward: location privacy and location awareness," in *SPRINGL'10: Proc. of the 3rd ACM Int'l Workshop on Security and Privacy in GIS and LBS*, 2010, pp. 1–3.

[10] 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.

[11] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *MobiSys'03: Proc. of the 1st Int'l Conf. on Mobile Systems, Applications and Services*, 2003, pp. 31–42.

[12] S. Guha and J. Birnholtz, "Can you see me now?: location, visibility and the management of impressions on foursquare," in *MobileHCI'13: Proc. of the 15th Int'l Conf. on Human Computer Interaction with Mobile Devices and Services*, 2013, pp. 183–192.

[13] 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.

[14] B. Henne, C. Kater, M. Smith, and M. Brenner, "Selective cloaking: Need-to-know for location-based apps," in *PST'13: Proc. of the 11th Annual Int'l Conf. on Privacy, Security and Trust*, 2013, pp. 19–26.

[15] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in GPS traces via uncertainty-aware path cloaking," in *CCS'07: Proc. of the 14th ACM Conf. on Computer and Communications Security*, 2007, pp. 161–171.

[16] 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.

[17] T. Jiang, H. J. Wang, and Y.-C. Hu, "Preserving location privacy in wireless LANs," in *MobiSys'07: Proc. of the 5th Int'l Conf. on Mobile Systems, Applications and Services*, 2007, pp. 246–257.

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

[19] J. Krumm, "Inference attacks on location tracks," in *Pervasive'07: Proc. of the 5th Int'l Conf. on Pervasive Computing*, 2007, pp. 127–143.

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

[21] 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 *CHI'11: Proc. of the 21st ACM Conf. on Human Factors in Computing Systems*, 2011, pp. 2409–2418.

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

[23] K. Micinski, P. Phelps, and J. S. Foster, "An empirical study of location truncation on android," in *MoST'13: Proc. of Mobile Security Technologies*, 2013.

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

[25] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREC'10: Proc. of the 2010 Int'l Conf. on Language Resources and Evaluation*, 2010.

[26] S. Patil, Y. Le Gall, A. J. Lee, and A. Kapadia, "My privacy policy: exploring end-user specification of free-form location access rules," in *USEC'12: Proc. of the 2012 Workshop on Usable Security*, 2012, pp. 86–97.

[27] S. Patil, G. Norcie, A. Kapadia, and A. J. Lee, "Reasons, rewards, regrets: Privacy considerations in location sharing as an interactive practice," in *SOUPS'12: Proc. of the 8th Symp. on Usable Privacy and Security*, 2012, pp. 5:1–5:15.

[28] S. Patil, G. Norcie, A. Kapadia, and A. Lee, "Check out where i am!: location-sharing motivations, preferences, and practices," in *CHI'12: Proc. of the 22nd ACM Conf. on Human Factors in Computing Systems (Extended Abstracts)*, 2012, pp. 1997–2002.

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

[30] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *SP'11: Proc. of the 2011 IEEE Symp. on Security and Privacy*, 2011, pp. 247–262.

[31] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. Le Boudec, "Protecting location privacy: Optimal strategy against localization attacks," in *CCS'12: Proc. of the 19th ACM Conf. on Computer and Communications Security*, 2012, pp. 617–627.

[32] K. P. Tang, J. I. Hong, and D. P. Siewiorek, "Understanding how visual representations of location feeds affect end-user privacy concerns." in *UbiComp'11: Proc. of the 13th ACM Int'l Conf. on Ubiquitous Computing*, 2011, pp. 207–216.

[33] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing twitter" big data" for automatic emotion identification," in *PASSAT'12: Proc. of the Int'l Conf. on Privacy, Security, Risk and Trust*, 2012, pp. 587–592.

[34] Y. Wang and I. H. Witten, "Inducing model trees for continuous classes," in *ECML'97: Proc. of the 9th European Conf. on Machine Learning*, 1997, pp. 128–137.

[35] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *MobiCom'11: Proc. of the 17th Annual ACM Int'l Conf. on Mobile Computing and Networking*, 2011, pp. 145–156.

[36] 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.