A Privacy-Aware Remapping Mechanism for Location Data

Guilherme Duarte
CISUC, CRACS/INESCTEC, and Department of Computer Science Faculty of Sciences, University of Porto, Porto, Portugal
guilherme.duarte@fc.up.pt

Mariana Cunha
CISUC, CRACS/INESCTEC, and Department of Computer Science Faculty of Sciences, University of Porto, Porto, Portugal
mariana.cunha@fc.up.pt

João P. Vilela
CISUC, CRACS/INESCTEC, and Department of Computer Science Faculty of Sciences, University of Porto, Porto, Portugal
jvilela@fc.up.pt

ABSTRACT

In an era dominated by Location-Based Services (LBSs), the concern of preserving location privacy has emerged as a critical challenge. To address this, Location Privacy-Preserving Mechanisms (LPPMs) were proposed, in where an obfuscated version of the exact user location is reported instead. Adding to noise-based mechanisms, location discretization, the process of transforming continuous location data into discrete representations, is relevant for the efficient storage of data, simplifying the process of manipulating the information in a digital system and reducing the computational overhead. Apart from enabling a more efficient data storage and processing, location discretization can also be performed with privacy requirements, so as to ensure discretization while providing privacy benefits. In this work, we propose a Privacy-Aware Remapping mechanism that is able to improve the privacy level attained by Geo-Indistinguishability through a tailored pre-processing discretization step. The proposed remapping technique is capable of reducing the re-identification risk of locations under Geo-Indistinguishability, with limited impact on quality loss.

CCS CONCEPTS

• **Security and privacy → Privacy protections; Usability in security and privacy**; • **Human-centered computing** → *Mobile devices.*

KEYWORDS

Location Privacy, Differential Privacy, Remapping, Location-Based Services, Data Discretization

ACM Reference Format:
Guilherme Duarte, Mariana Cunha, and João P. Vilela. 2024. A Privacy-Aware Remapping Mechanism for Location Data. In *The 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24), April 8–12, 2024, Avila, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3605098.3636050

1 INTRODUCTION

Location-Based Services (LBSs), while undeniably valuable in enhancing the convenience and efficiency of our daily lives, can give rise to significant privacy concerns. These services rely on tracking and storing an individual's real-time location data, which, if misused or mishandled, could lead to severe repercussions. From the perspective of personal privacy, the constant monitoring of an individual's movements can paint an intimate picture of their habits, preferences, and even sensitive activities [9, 12, 24]. In the wrong hands, this data could be exploited for targeted advertising, identity theft, or surveillance, compromising individuals' autonomy and security. Additionally, under inference attacks, location-based services might (un)intentionally disclose sensitive locations, like home, workplace, health and religious institutions, as well as information about users, their habits and conditions, thus making users vulnerable to potential threats.

Over the past two decades, important achievements have been accomplished in user protection, specifically in the fields of anonymization and obfuscation techniques [6, 13]. Anonymization involves modifying or removing personally identifiable information from datasets, making it challenging to link specific data points to individual users. On the other hand, obfuscation mechanisms introduce noise or perturbations to location data, making it more challenging to pinpoint an individual's exact location. Despite their differences, both methods make use of spatiotemporal generalization, which involves aggregating or reducing the granularity of location data to a certain level. By doing so, they mask precise details about a user's movements and activities, preserving their anonymity while still providing useful information for analysis or services.

Geo-Indistinguishability [2] is acknowledged as the of state of the art in Location Privacy-Preserving Mechanisms (LPPMs), where Planar Laplace (PL) was the first mechanism proposed to achieve it. Built upon the foundation of differential privacy [7], Geo-Indistinguishability ensures that even when sharing location information for various services or applications, an individual's true whereabouts remains hidden within a radius r with a level of privacy that depends on the radius, while providing a delicate balance between utility and privacy.

Remapping techniques have been proposed for Planar Laplace to increase the utility of the queries without degrading the privacy level by feeding the noisy generated location to a remapping function which relocates them in a more suitably new location, considering the remapping function metric [4]. In fact, the PL mechanism with optimal remapping is considered the state-of-the-art of Geo-Indistinguishability in sporadic location privacy [21]. The optimal remapping techniques only use the current obfuscated LPPM output and the mobility profile of the user for mapping an obfuscated location into a grid-based discrete location [2, 4]. Remapping targets preserving the privacy guarantees while maintaining the quality loss in the same order of the noised location generated through the privacy-preserving mechanism. However, we verified that after applying a privacy-preserving mechanism and remapping into a grid, the users' locations become much more unique due to the

SAC '24, April 8–12, 2024, Avila, Spain

$\mathcal{X}$	Set of true locations
$\mathcal{X}^*$	Set of discrete true locations
$\mathcal{Z}$	Set of obfuscated locations
$\mathcal{Z}^*$	Set of discrete obfuscated locations
ϵ	Privacy parameter
$\mathcal{K}(\cdot)$	Obfuscation mechanism
$\mathcal{K}_\epsilon(\cdot)$	Obfuscation mechanism with a fixed ϵ
$PL_\epsilon(\cdot)$	Planar Laplace Function
$C_\epsilon(\cdot)$	CDF for a radius of obfuscation
r	Radius of obfuscation
$\mathcal{G}$	Set of cells from a grid discretization
$G(\cdot)$	Grid Discretization Function
s	Constant cell spacing on a grid

Table 1: General Notations

spreading effect, which could pose a threat if the anonymization poorly protects user traits.

Therefore, we propose a novel algorithm which produces a remapping function from the cells of a uniform grid into itself. Our proposed remapping mechanism aggregates groups of cells from a uniform partitioning of the domain of locations. This process of cell aggregation takes into consideration the frequency of locations that actually appear in the dataset, to produce a smaller set of utilized cells, i.e. cells effectively used during the discretization process. At a small cost in terms of utility, our discretization method makes it more difficult for an adversary to re-identify individuals. Our remapping function strategically transforms cells within the grid, effectively diminishing the impact of weighted noise introduced through Geo-Indistinguishability techniques, yielding a more sensible remapping of locations, ensuring that the privacy enhancements gained from obfuscation methods are not compromised, whilst resulting in a discrete dataset.

The major contributions of this paper are summarized as follows:

- identification of flaws from remapping Laplacian noised locations into a uniform grid-based discretization mechanism;
- proposal of a novel algorithm which addresses the flaws identified from a uniform grid at a small cost of utility.

The remainder of the paper is organized as follows: we review the existing works in Section 2. Section 3 describes our remapping mechanism. We evaluate our proposed method in Section 4 and discuss the obtained results. Section 5 concludes our work. The notation used throughout the paper is presented in Table 1.

2 BACKGROUND AND STATE OF THE ART

This section provides an overview of background concepts and state-of-the-art approaches that are taken into consideration in the development of the Privacy-Aware Remapping mechanism. First, in Section 2.1, we grasp the essence of what is a mechanism that achieves Geo-Indistinguishability and consider one such method in Section 2.2. In Section 2.3 we discuss about discretization techniques and their properties. At last, we introduce the main focus of this work: the composition of obfuscation methods alongside discretization techniques, in Section 2.4.

2.1 Geo-Indistinguishability

Geo-Indistinguishability [2] is a privacy concept and technique that ensures that an individual's exact location is indistinguishable from a set of nearby locations, thereby preventing the precise identification of a user's movements and activities. Geo-Indistinguishability achieves this by introducing noise or perturbations to the location data, making it challenging to link specific location traces to a particular individual. The level of indistinguishability is controlled by a parameter called the privacy budget ϵ, which determines the amount of noise added to the data. The main idea behind a Geo-Indistinguishable mechanism is the guarantee that the user location x is indistinguishable to any other nearby location x' based on the obfuscated report z.

Formally [17], denoting by $\mathcal{K}$ an obfuscation mechanism which assigns to every true location $x \in \mathcal{X}$ a probabilistic distribution on $\mathcal{Z}$, the set of all obfuscated locations, this mechanism satisfies ϵ-geo-indistinguishability iff for all $x, x' \in \mathcal{X}$:

$$d_{\mathcal{P}}(\mathcal{K}(x), \mathcal{K}(x')) \leq \epsilon d_{\mathrm{x}}(x, x') \tag{1}$$

where $d_{\mathrm{x}}(\cdot)$ is any distance function and $d_{\mathcal{P}}(\cdot)$ is the multiplicative distance between two distributions, defined as:

$$d_{\mathcal{P}}(\sigma_1, \sigma_2) = \sup_{S \in \mathcal{S}} \left| \log \frac{\sigma_1(S)}{\sigma_2(S)} \right| \tag{2}$$

where σ_1 and σ_2 are two distributions on some set $\mathcal{S}$, and following the convention that $\mathcal{L} = \left| \log \frac{\sigma_1(S)}{\sigma_2(S)} \right| = 0$ if $\sigma_1(S) = \sigma_2(S) = 0$ and $\mathcal{L} = \infty$ if only one of the two is 0.

Intuitively, condition (1) states that the probability of reporting location z while standing in location x is similar to that of standing in any location x' [16]. In particular, both probabilities differ at most by the distance between x and x' factored by a small constant ϵ. This constant is usually set to $\epsilon = l/r$ [2], which represents a simple way to specify a user's privacy requirements - level of privacy l within a radius r, enforcing that any x' within r distance of x discloses l information, at most.

2.2 Planar Laplace Mechanism

Planar Laplace (PL) [2] was the first mechanism proposed to satisfy Geo-Indistinguishability and consists of adding 2-dimension Laplacian noise centered at the true user location x. Formally, for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, the probability density function (pdf) is given by:

$$p_x(z) = \frac{\epsilon^2}{2\pi} e^{-\epsilon d(x,z)} \tag{3}$$

Obtaining z, the obfuscated location, from x can be efficiently done using polar coordinates [2]:

(1) draw θ uniformly in $[0, 2\pi)$
(2) draw ρ uniformly in [0,1)
(3) set $r = C_\epsilon^{-1}(\rho)$

where the cumulative densitity function (cdf) $C_\epsilon(r)$ represents the probability that the radius of the random generated point falls between 0 and r, which uses PL's cumulative distribution function defined in (3). Therefore, using the Lambert W function at branch -1, the inverse function is defined as:

$$C_\epsilon^{-1}(\rho) = -\frac{1}{\epsilon}(W_{-1}(\frac{\rho - 1}{e}) + 1) \tag{4}$$

Finally, simply report $z = x + \langle r\cos(\theta), r\sin(\theta)\rangle$. We will denote by $PL_\epsilon : X \to \mathcal{Z}$ the above mechanism.

2.3 Data Discretization

Location Data discretization is a crucial pre-processing step in managing and sharing location-based data effectively while upholding individual privacy. It involves transforming continuous and precise location coordinates in $\mathbb{R}^2$ into a discrete set of points $\mathbb{W}$, such as cells or clusters. Additionally, it allows us to reduce the granularity of the data, making it less precise and less likely to pinpoint individuals' exact locations, while still providing high levels of utility. Our proposed mechanism, Privacy-Aware Remapping, will dispatch locations to others and take advantage of the finite set of potential locations resulting from discretization.

One widely used discretization approach is a grid-map method, used in various scenarios [2, 4, 11, 18, 19, 22, 25–27], where geographic regions are partitioned into a grid of uniform cells. Another technique involves clustering methods [8, 10, 23, 29], which group similar data points together, creating clusters that represent geographical areas. Additionally, Voronoi diagrams [14, 20] divide a geographical space into cells based on the proximity to specific data points, ensuring that each region is associated with the nearest data point.

We decided to proceed with the study of the grid-based discretization technique, mentioned as the *uniform grid* throughout the paper, due to its distinct advantages. One of the key reasons for selecting this method is its simplicity in storage. The uniform grid discretization can be efficiently represented as a matrix, making it an ideal choice for data storage as well as data manipulation. Furthermore, this technique allows for precise control over the size of the discrete locations, which is especially valuable when analysing data transformations, like privacy-preserving methods such as Geo-Indistinguishability.

A uniform grid, denoted by $\mathcal{G}$, partitions geographic regions into cells of constant size. Grid discretization is a powerful yet simple technique used to efficiently represent and process location data. Given four corners of a *bounding box* as well as a constant value of a cell spacing, the mechanism divides the box in $n = n_0 \times n_1$ cells, represented by a n_0-by-n_1 matrix. Real-time services, such as navigation systems, and points-of-interest (POI's) finders, can benefit from this structured representation since the obfuscation can be calculated in real-time - simply pinpoint the cell it contains the ground-truth, i.e. the real location. Moreover, the quality loss resulting from this discretization is bounded and predictable since each cell maintains a constant size. Denoting by s the grid cell's spacing and considering that the reported location from a cell is its center, the maximum quality loss obtainable is $s/\sqrt{2}$ (the points within the cell furthest from its center are its corners).

We will denote the function which discretizes real locations to a cell of the grid $\mathcal{G}$ as $G : \mathcal{X}/\mathcal{Z} \to \mathcal{G}$.

2.4 Remapping Locations

We are now prepared to discuss how to add a layer of discretization after applying Laplacian noise to a real location, which is the main focus of this work.

In [2], it has been proven that grid map discretization preserves Geo-Indistinguishability. The authors have defined the probabilistic mechanism $K_\epsilon : \mathcal{G} \to P(\mathcal{G})$, where $K_\epsilon(c|x)$ represents the probability of reporting the cell c when the actual ground-truth location is x, with $x \in \mathcal{X}$ and $c \in \mathcal{G}$. The mechanism involves generating (r, θ) and computing z, as previously described, and then remapping z to the closest point c on $\mathcal{G}$. Formally, given $x \in \mathcal{X}$ we obtain $c = (G \circ PL_\epsilon)(x)$ and report its center.

Let us focus on Figure 1. We extracted a user's trajectory from the data we used throughout the paper and applied PL_ϵ to demonstrate how it affects the utilized cells (i.e. cells from the grid that contain location reports). As it can be seen, the application of a mechanism such as PL spreads out the true locations, sometimes even to cells from the grid which were previously empty. This result implies that a larger amount of discrete points (center of cells) are needed to discretize every true location, i.e. the number of unique locations in the discrete set of true locations, $\mathcal{X}^*$, will be much smaller than the number of unique locations in the discrete set of obfuscated locations, $\mathcal{Z}^*$.

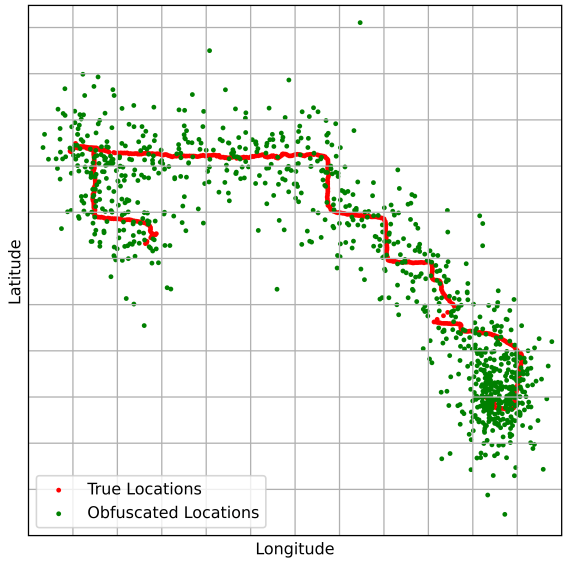

Figure 1: Noise added by PL mechanism and how it affects the utilized cells

This result can contribute to privacy threats to the individual, as we will see in Section 4.2. The remapping that our algorithm produces, introduced in the following section, will come in as a privacy context-aware choice which decreases densely the amount of clusters needed to discretize the data, while still achieving Geo-Indistinguishability. Additionally, we will verify how benefiting the use of a smaller set of clusters will also dramatically decrease the number of individuals affected by re-identification attacks thought their most visited and preferred locations.

3 A PRIVACY-AWARE REMAPPING MECHANISM

The algorithm we propose generates a remapping function, denoted by $\mathcal{R}$, which maps individual cells within a grid to themselves. The distinctive property that $\mathcal{R}$ offers is its ability to minimize the weighted quality loss inherent from the conjugation of Laplacian noise after remapping to a grid. When applied to a specific cell, $\mathcal{R}(c)$ will identify and designate the cell that minimizes the weighted quality loss, considering all the possible cells which PL_ϵ might report locations from the cell c to. We will then verify how $\mathcal{R}$ is actually nor injective, i.e. given two cells $c_1, c_2 \in \mathcal{G}$ s.t. $c_1 \neq c_2$ does not imply that $\mathcal{R}(c_1) \neq \mathcal{R}(c_2)$; nor surjective, i.e. not for all c_2 must exist a c_1 s.t. $\mathcal{R}(c_1) = c_2$. So the number of cells which will get reported with $\mathcal{R}$ will be smaller than the number of cells that a uniform grid requires. We will show how that provides extra privacy guarantees, thus making re-identification of users more challenging. Furthermore, this remapping generates a grid composed of multiple aggregated cells, each contained by unique cells and a single output center which minimizes our metric, so it is able to determine which areas will need more or less generalization.

This section is divided as follows: in Section 3.1 we present the pseudocode to the generator of the Privacy-Aware Remapping; in Section 3.2 we explain how remapping was designed to be used and some of its properties; and finally, we discuss the computation complexity in Section 3.3.

3.1 Algorithm in a Nutshell

The pseudocode of the generation of the remapping function is listed in Algorithm 1. The inputs to the algorithm are the dataset $\mathbb{M}$, a uniform grid $\mathcal{G}$, and the obfuscation radius r related to the Laplacian noise. The algorithm will then return a remapping map $\mathcal{R} : \mathcal{G} \to \mathcal{G}$.

In the initialization phase, the remap $\mathcal{R}$ is initialized as an empty map (line 1). Afterwards, using the subroutine *generateCellsWeight* (line 2), we build a weight function w s.t. $w(c)$ holds the number of ground-truth reports which lay on the grid cell c. This can be achieved by a simple iteration over every report from the dataset, translating the coordinates into cells from the grid, and increment accordingly.

The algorithm then enters the main loop (lines 3-17), which iterates over all cells c from $\mathcal{G}$ and calculates the best candidate c' to get reported instead, according to a metric we will now describe. Let $D_r(c)$ denote the set of cells where Planar Laplace might send locations in c to, i.e. $D_r(c)$ will contain every cell c' such that the distance from c to c' is less than or equal to r. In the *for* conditions of lines 6 and 8, we verify if c' and c'' are in $D_r(c)$, respectively, where d_{gcd} denotes the great circle distance.

Finally, based on the *Bayesian* remap [4], we define the optimal candidate for remapping c as the cell that effectively minimizes the weighted generated quality loss, formally described as:

$$R(c) = \underset{c' \in D_r(c)}{\arg\min} \sum_{c'' \in D_r(C)} w(c'') \cdot d_e(c', c'') \tag{5}$$

where $d_e(\cdot)$ denotes the euclidean distance between the two cells, since each cell can be represented as i and j offsets from matrix of the grid $\mathcal{G}$. This is accomplished on lines 8-14.

Algorithm 1 Pseudocode for the generator of $\mathcal{R}$

Input: Dataset $\mathbb{M}$
Input: Uniform grid $\mathcal{G}$
Input: Obfuscation radius r
Output: Cell remapping $\mathcal{R}$

1: $\mathcal{R} \leftarrow$ *initialize map*
2: $w \leftarrow generateCellsWeight(\mathbb{M}, \mathcal{G})$
3: **for** $c \in \mathcal{G}$ **do**
4: $\quad nc \leftarrow c$
5: $\quad lowestError \leftarrow \infty$
6: $\quad$ **for** $c' \in \mathcal{G}$ s.t. $d_{gcd}(c, c') \leq r$ **do**
7: $\quad\quad error \leftarrow 0$
8: $\quad\quad$ **for** $c'' \in \mathcal{G}$ s.t. $d_{gcd}(c, c'') \leq r$ **do**
9: $\quad\quad\quad error \leftarrow error + w(c'') \cdot d_e(c', c'')$
10: $\quad\quad$ **end for**
11: $\quad\quad$ **if** $error < lowestError$ **then**
12: $\quad\quad\quad nc \leftarrow c'$
13: $\quad\quad\quad lowestError \leftarrow error$
14: $\quad\quad$ **end if**
15: $\quad$ **end for**
16: $\quad \mathcal{R}(c) = nc$
17: **end for**

3.2 Remapping Planar Laplace Points Using $\mathcal{R}$

Using our Privacy-Aware Remapping $\mathcal{R}$ will result in an extra step before reporting the obfuscated location. As before, given the actual location $x \in \mathcal{X}$, we generate (r, θ) and compute $z \in \mathcal{Z}$ following the PL methodology (Section 2.2). Afterwards, we get $c \in \mathcal{G}$, the cell where z is contained on the uniform grid and we finally report the cell $\mathcal{R}(c)$. Formally, we obtain $c = \mathcal{R}(G(PL_\epsilon(x)))$ and report its center.

Let us discuss some properties which this remapping provides. The computation of the function $\mathcal{R}$ will generate group of cells which will report the same cell. Let H_c represent the set of cells which get remapped to c, i.e. $\forall c' \in \mathcal{G}$:

$$c' \in H_c \text{ iff } \mathcal{R}(c') = c \tag{6}$$

Note that c might not necessarily belong to H_c: the algorithm might find a more suitable cell from the set $D_r(c)$ which decreases the overall weighed quality loss (Equation 5). This abstraction of the remapping $\mathcal{R}$ allows us to see the uniform-grid as a coarser grid, by considering every set $H_c \neq \emptyset$ as unified cells which have as its center the same cell c.

Additionally, due to the nature of the algorithm, we found that it is highly likely that there will always exist some cell c such that $|H_c| > 1$. That can be justified by the simple fact that two neighbour cells c_1, c_2 will have identical sets of obfuscation centered at each cell, respectively $D_r(c_1)$ and $D_r(c_2)$, so the one cell which minimizes the weighed quality loss of each cell will be, most likely, one cell on the intersection of both sets. Naturally, this result depends on the grid spacing s, the privacy parameter ϵ, as well as the obfuscation radius r, since these will determine how much the obfuscation sets of neighboring cells will intercept each other. Therefore, assuming an appropriate configuration of the parameters according to the dataset and the desired utility/privacy trade-off,

the $\mathcal{R}$ remapping is, by definition, not injective, implying that is also not surjective, since the domain and co-domain sets are the same, then not for all c_2 must exist a c_1 s.t. $\mathcal{R}(c_1) = c_2$, through the pigeonhole principle.

3.3 Computational Complexity

In order to determine the exact computational complexity of Algorithm 1, let us consider a dataset with $|\mathbb{M}|$ reports, a grid $\mathcal{G}$ with n cells of constant size s, as well as a radius of obfuscation r.

At the initialization phase, the first instruction (line 1) takes constant time. Computing the weight w of each cell (line 2) requires to verify in what cell each report from $\mathbb{M}$ lays. That can be done in $O(|\mathbb{M}|)$.

The main loop from line 6 to 15, which runs n times, requires some enhancements to avoid an $O(n^3)$ algorithm, which would quickly become infeasible for grids with tens of thousands of cells. For each $c \in \mathcal{G}$, we are interested to perform a nested-loop on every cell in $D_r(c)$, which are all the cells that are at a distance of at most r from c, as explained before. Instead of linear search between every cell to verify if the condition is met (as described in Algorithm 1), one can simply consider a square of cells centered at c composed by $\lceil 2 \cdot \frac{r}{s} + 1 \rceil^2$ cells, since the obfuscation circle with center c and radius r is inscribed in this square, as it can be seen in Figure 2 (the green area corresponds to $D_r(c)$ and the red area to the search space of the algorithm). So the complexity of this loop is $O(n \cdot \lceil \frac{r}{s} + 1 \rceil^4)$, which can be simplified to $O(n \cdot (\frac{r}{s})^4)$ assuming r is rounded to the next closest multiple of s, as well as following the properties of the Big-O.

Therefore, the overall complexity of our algorithm is $O(|\mathbb{M}| + n \cdot (\frac{r}{s})^4)$, so it grows as the grid spacing decreases ($\frac{r}{s}$ and n increases) as well as when considering a greater obfuscation radius. Note that the main loop can be done concurrently, so it is possible to divide the workload among multiple processing units or threads, resulting in a feasible execution time when considering large datasets and small levels of granularity.

In the next section, we provide reasoning for the additional computational cost needed to construct the remapping function, as opposed to the straightforward use of the uniform grid.

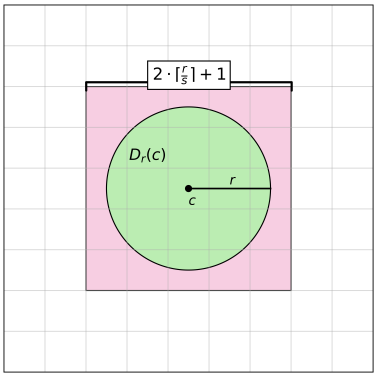

Figure 2: Circle of obfuscation in green and algorithm's square of search in red

4 EVALUATION AND DISCUSSION

To evaluate the effectiveness of the conjugation of PL mechanism with remapping using only the uniform grid and the additional remapping of $\mathcal{R}$, we selected a real mobility dataset, Geolife [31], which was collected in a period of over three years from GPS devices. The dataset contains data from 182 users, 17,621 trajectories and roughly 25 million reports. Following [15], we first limited the distribution of reports to a bounding box over 5$^{\text{th}}$ ring road of Beijing, China. It is defined from South and North by the latitudes 39.753, 40.026, and from West and East by longitudes 116.199, 116.547, still leaving us with approximately 16 million reports. This division allowed us to focus on a high traffic urban area surrounded by the suburbs with a lower density.

As constant spacing of the cell's grid, we fixed the value of 100 meters [3, 28], which we found to provide a reasonably high level of resolution for most practical purposes. For values of the privacy budget ϵ, we used multiple values in the typical ranges of privacy-preserving mechanisms for continuous reports [1, 5, 15], specifically $\epsilon = [4, 8, 16, 32]$ km^{-1}. For the PL, this corresponds to an average obfuscation of $[500, 250, 125, 62.5]$ m, respectively. With these values across the 100 meter intervals of the grid cells, we can investigate scenarios where the obfuscation, on average, extends far beyond the confines of the location's cell, approaches the cell boundary, or remains entirely contained within the cell itself.

For the value of r, the obfuscation radius, let us recall (Equation 4) the inverse cumulative density function of PL and how this function, given a probability ρ, returns the radius r for which the probability of falling within that radius is equal to ρ. Note that:

$$\lim_{\rho \to 1} C_\epsilon^{-1}(\rho) = -\frac{1}{\epsilon}(-\infty + 1) = +\infty \tag{7}$$

since $\lim_{\rho \to 1} \frac{\rho - 1}{\epsilon} = 0$ and $\lim_{z \to 0} W_k(z) = -\infty$ for all $k \neq 0$, so the obfuscation radius given by C_ϵ^{-1} could get as large as possible as ρ gets closer to 1. Therefore, there is not a ρ we can feed to C_ϵ^{-1} which would bound all the noise added from all obfuscated locations. So we decided to set $r = C_\epsilon^{-1}(\rho) + s/\sqrt{2}$, with ρ equal to 95%. The additional $s/\sqrt{2}$ represents the maximum quality loss generated by discretizing a real location on a uniform grid, as mentioned in Section 2.3. This way, we can focus on the obfuscation circle centered at each cell $c \in \mathcal{G}$ with radius r, knowing that the 95$^{\text{th}}$ percentile of all obfuscated locations at c are contained in the circle.

4.1 Quality Loss

Quality loss is a point-by-point metric, measuring the quality lost between the ground-truth and the data obtained after applying a privacy-preserving mechanism, i.e. for an original location $x \in \mathcal{X}$ and the respective obfuscated location $z \in \mathcal{Z}$, the quality loss is given by $d_e(x, z)$, the Euclidean distance metric. When we are considering remapping after using PL, we actually compute $d_e(x, c)$ where $c \in \mathcal{G}$ represents the center of the cell where z is contained.

Figure 3 quantifies the quality loss of our Privacy-Aware Remapping compared to a plain Uniform Remapping and Planar Laplace without Remapping (PL_ϵ), for different values of privacy budget, ϵ. Initially, one can see how little affects remapping PL to a uniform grid. Since the maximum introduced error by this operation is $s/\sqrt{2}$, this cost will be negligible as ϵ decreases, i.e. as obfuscation

increases. The weighed quality loss used in our proposal (see Equation 5), although related, does not necessarily decrease the overall quality loss. Let $x \in \mathcal{X}$, z the output of PL_ϵ evaluated on x, and c the cell where z is contained. With high probability (95% from the configuration of the obfuscation radius r), the true location x will be contained in the circle of obfuscation centered at cell c with radius r. On the other hand, $c' = \mathcal{R}(c)$, by definition, will be also a cell from the circle. Therefore, the overall introduced error must be bounded by the diameter of the circle, i.e. $d_e(x, c') \leq 2 \cdot r$.

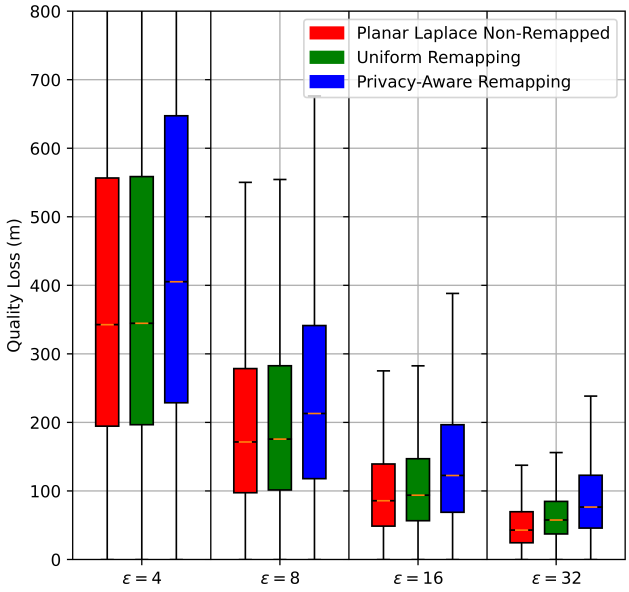

Figure 3: Quality loss obtained when using Planar Laplace with no Remapping and when applying Uniform and Privacy-Aware Remapping, for different ϵ's

Additionally, we found that the smaller the ϵ, the larger the loss of utility with remapping is. This behaviour is already expected when applying PL, i.e. the average obfuscation from a privacy budget is given by $2/\epsilon$ so the obfuscation amount increases inversely to ϵ. As it can be observed, using the Privacy-Aware Remapping introduces an average obfuscation of $2/\epsilon + k$, where $k > 0$ also grows inversely to ϵ but $k \ll 2/\epsilon$. For instance, an epsilon value of 4 km^{-1} resulted in an average obfuscation of 414 meters and, consequently, to an average quality loss of 416 and 478 meters when applying Uniform Remapping and Privacy-Aware Remapping, respectively, which is still under the expected obfuscation of 500 meters. Therefore, we argue that the amount of noise generated with $\mathcal{R}$ remapping is in the same order of the noise from Uniform Remapping.

In the upcoming sections, we will furnish beneficial outcomes that can rationalize the introduced supplementary error.

4.2 Number of utilized cells

We focused on comparing the number of utilized cells with and without the $\mathcal{R}$ remapping, having as baseline the number of unique cells needed to discretize the non-obfuscated data, which we refer as ground-truth. As previously described, an utilized cell is a cell from the grid which has location reports contained. Table 2 summarizes the obtained results. Firstly, note that the number of utilized cells without obfuscation is constant to all privacy parameters ϵ, equal to 41120. Following, note the effect of the ϵ's values have on the number of utilized cells. As mentioned in Section 2.4, applying PL_ϵ will spread out the true locations (see Figure 1) and, as the privacy parameter ϵ increases, so does the radius of obfuscation. Therefore, Uniform Remapping contributed to a large increase of the size of this set and, consequently, there will be more unique discrete locations than unique ground-truth locations, leading to user's trajectories to become overall more unique. This result can lead to some privacy threats in the case where the anonymity mechanism performs poorly, for instance when the applied noise maps true locations into non-admissible ones or sparsely frequented, like into the sea or mountain. Although possible, it is unlikely to receive queries at such locations, so an adversary can take advantage of this information.

	Number of utilized cells			
	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$	$\epsilon = 32$
Ground-Truth Remapped	41120			
Uniform Remapping	81945	74191	65367	56113
Privacy-Aware Remapping	29255	24447	21327	19539

Table 2: Number of utilized cells when using Ground-Truth Remapped Data and when applying Uniform and Privacy-Aware Remapping for different ϵ's

On the other hand, our Privacy-Aware Remapping was able to reduce around 300% of the utilized cells in comparison with Uniform Remapping. At a small cost on utility, the proposed remapping drastically reduces the number of discrete points, which has its advantages at the computational level, also leading to better results under inference attacks, as we will see in the next section. Furthermore, the remapping into non-admissible locations will rarely happen, since those locations have no weight associated, not contributing to the weighted quality loss metric used to determine the optimal $\mathcal{R}(c)$, for every c.

4.3 Top-N Re-Identification Attack

The top-N re-identification attack [30] measures the risk of a privacy threat that revolves around the concept of identifying individuals based on their location data. In this attack, an individual is considered re-identifiable if their top-N visited locations are unique, even if the actual identity of the person is anonymized. This attack leverages the uniqueness of an individual's movement patterns and frequently visited places to de-anonymize them.

For each N, we compute the anonymity sets, namely the number of users with the same top-N preferential locations. Therefore, we define a user as *re-identifiable* if the size of the anonymity set is 1, i.e. the user's top-N most visited locations identify the user without ambiguity, meaning that there is no other user with the same top-N visited locations. If a noise-based privacy mechanism, like PL_ϵ is applied to the data, we add to the condition of being considered as re-identified the following: the user's top-N produced from the unaltered data must be the same as the top-N produced

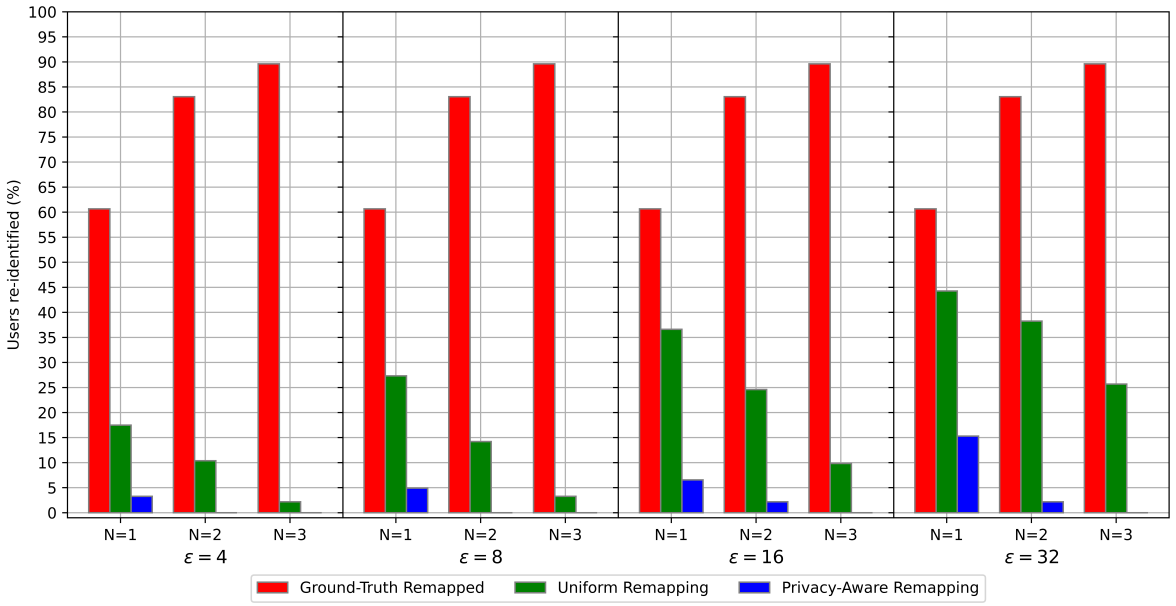

Figure 4: Percentage of re-identified users with the Top-N Re-Identification Attack, i.e. the real locations remapped to the uniform grid, and for the Uniform and Privacy-Aware Remapping, for different privacy parameters ϵ's and N's

from the obfuscated data. This way, even if a user is in a unitary-sized anonymity set, if the obfuscation mechanism changed the top-locations, then we consider the user as non re-identifiable, although, in this case, the user can be considered as unique.

Figure 4 depicts the percentage of re-identified users when applying the top-N attack with the ground-truth data, i.e. the non-obfuscated data remapped into a grid, and using the Uniform and Privacy-Aware Remapping. Note that the results from the non-obfuscated data do not vary with different values of the privacy parameter. Following [30], we considered the top-1, top-2 and top-3 locations of each user. Intuitively, for the ground-truth data, top-1 is included in top-2, as well as top-2 in top-3, so one can only expect a greater amount of re-identifications as N increases. As it can be seen at the ground-truth results, at a such fine level of granularity (regions of 100 meters), an attacker can easily re-identify a large chunk of users, even when considering small sets of top-locations, reaching upwards to 90% of re-identification when considering the top-3 locations of each user.

Due to the extra rule we added to consider a user as re-identified in case a privacy mechanism is applied, re-identification no longer grows as the set of top-locations enlarges. As the value of N increases, most likely the user's top-N produced from the unaltered data differs from the top-N produced from the obfuscated data. Notice now how decreasing the privacy parameter, where one should expect privacy to be favoured, also decreases the overall re-identification. As ϵ decreases, higher amounts of obfuscation gets added to the real location, so it becomes easier to protect a user's top-locations. For example, for $\epsilon = 8$ km^{-1} and $N = 1$, the top-N attack with Uniform Remapping was able to re-identify 27% in comparison to 83% when not applying any obfuscation to the data, which are already substantial results. Privacy-Aware Remapping was able to decrease this value even further to around 5%. This represents an improvement of 81% in comparison to the Uniform Remapping and 94% to the ground-truth data. In a lot of other cases, our mechanism was able to achieve total user protection, representing considerable results in comparison to the Uniform Remapping.

Finally, our Privacy-Aware Remapping decreases immensely the re-identification, most of the times achieving total user protection, thus showing an adequate utility/privacy trade-off, since a quality loss of the same order as the Uniform Remapping is accompanied by a large reduction of the number of re-identified user.

5 CONCLUSIONS

As Location-Based Services become ubiquitous, the need for effective privacy mechanisms cannot be overstated. State-of-the-art approaches resolve into Geo-Indistinguishability, a privacy concept that prevents the precise identification of a user's location. For an extra layer of protection, discretization comes into play, transforming continuous locations into a discrete set of points. We have verified how the conjugation of these methods, namely Planar Laplace with a grid-based discretization technique (mentioned as Uniform Remapping in this paper), highly increases the number of utilized cells, i.e. the cells needed to fully discretize a dataset. The sparseness of locations resulting from the large number of utilized cells, can lead to threats coming from non-admissible locations, thus affecting the performance of anonymization mechanisms.

In response to this challenge, we have introduced a novel approach: Privacy-Aware Remapping. This mechanism builds upon the foundation of a grid-based discretization technique and unifies cells to enhance privacy protection. While it introduces a marginal

amount of extra quality loss, this loss is directly correlated with the chosen privacy parameter in PL_ϵ, therefore the Privacy-Aware Remapping introduces noise in the same order as Uniform Remapping does.

One of the key strengths of our proposed mechanism lies in its significant reduction of cluster regions, referred to as the utilized cells. Our remapping function, by design, is not surjective, resulting in a smaller number of output cells compared to the total available cells. This reduction not only streamlines data storage and processing but also minimizes the risk of potential breaches.

Privacy-Aware Remapping demonstrates a robust defense against re-identification attacks, particularly those leveraging top-N locations. In most of the cases, it is able to achieve a level of protection tantamount to total user anonymity. This outcome reinforces the viability and effectiveness of our approach in securing the privacy of individuals utilizing location-based services.

For a future work, it would be interesting to evaluate the behaviour of the proposed remapping technique against attacks other than the top-N. As the number of discrete regions tend to decrease significantly, more information about a user tends to be aggregated among others, and, therefore, we expect that this mechanism behaves competently facing other inference attacks.

ACKNOWLEDGMENTS

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020 and project CISUC-UID/CEC/0 0326/2020 and by European Social Fund, through the Regional Operational Program Centro 2020. This work was performed in the scope of the Smart Networks and Services Joint Undertaking (SNS JU) under the EU Horizon Europe programme PRIVATEER under Grant Agreement No. 101096110. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the EU or SNS JU. Mariana Cunha wishes to acknowledge financial support by the Portuguese funding institution Fundação para a Ciência e a Tecnologia (FCT) under the grant 2020.04714.BD.

REFERENCES

[1] Raed Al-Dhubhani and Jonathan M. Cazalas. 2018. An adaptive geo-indistinguishability mechanism for continuous LBS queries. *Wireless Networks* 24, 8 (01 Nov 2018), 3221–3239. https://doi.org/10.1007/s11276-017-1534-x

[2] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-Indistinguishability: Differential Privacy for Location-Based Systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security* (Berlin, Germany) *(CCS '13)*. Association for Computing Machinery, New York, NY, USA, 901–914.

[3] Filipe Batista e Silva, Javier Gallego, and Carlo Lavalle. 2013. A high-resolution population grid map for Europe. *Journal of Maps* 9, 1 (2013), 16–28.

[4] Konstantinos Chatzikokolakis, Ehab Elsalamouny, and Catuscia Palamidessi. 2017. Efficient utility improvement for location privacy. *Proceedings on Privacy Enhancing Technologies* 2017, 4 (2017), 308–328.

[5] M. Cunha, R. Mendes, and J. P. Vilela. 2019. Clustering Geo-Indistinguishability for Privacy of Continuous Location Traces. In *4th International Conference on Computing, Communications and Security (ICCCS)*. IEEE, 1–8.

[6] Mariana Cunha, Ricardo Mendes, and João P Vilela. 2021. A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer science review* 41 (2021), 100403.

[7] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation*, Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–19.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, Vol. 96. 226–231.

[9] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2010. Show Me How You Move and I Will Tell You Who You Are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS* (San Jose, California) *(SPRINGL '10)*. Association for Computing Machinery, New York, NY, USA, 34–41. https://doi.org/10.1145/1868470.1868479

[10] Marco Gramaglia, Marco Fiore, Alberto Tarable, and Albert Banchs. 2017. $k^{\tau,\epsilon}$-anonymity: Towards Privacy-Preserving Publishing of Spatiotemporal Trajectory Data. arXiv:1701.02243 [cs.CY]

[11] Yuzhou Jiang, Emre Yilmaz, and Erman Ayday. 2023. Robust Fingerprint of Location Trajectories Under Differential Privacy. In *Proceedings on Privacy Enhancing Technologies. Privacy Enhancing Technologies Symposium*, Vol. 2023. NIH Public Access, 5.

[12] John Krumm. 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing* 13, 6 (2009), 391–399.

[13] Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, and Yong Xiang. 2018. Location privacy and its applications: A systematic study. *IEEE access* 6 (2018), 17606–17624.

[14] Chunguang Ma, Changli Zhou, and Songtao Yang. 2015. A voronoi-based location privacy-preserving method for continuous query in LBS. *International Journal of Distributed Sensor Networks* 11, 3 (2015), 326953.

[15] Ricardo Mendes, Mariana Cunha, and João P. Vilela. 2020. Impact of Frequency of Location Reports on the Privacy Level of Geo-indistinguishability. *Proceedings on Privacy Enhancing Technologies* 2020, 2 (2020), 379 – 396.

[16] Ricardo Mendes, Mariana Cunha, and João P Vilela. 2023. Velocity-Aware Geo-Indistinguishability. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*. 141–152.

[17] Ricardo Mendes and João Vilela. 2018. On the Effect of Update Frequency on Geo-Indistinguishability of Mobility Traces. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks* (Stockholm, Sweden) *(WiSec '18)*. Association for Computing Machinery, New York, NY, USA, 271–276.

[18] Mehmet Ercan Nergiz, Maurizio Atzori, and Yucel Saygin. 2008. Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*. 52–61.

[19] Kun Niu, Huiyang Zhang, Tong Zhou, Cheng Cheng, and Chao Wang. 2019. A novel spatio-temporal model for city-scale traffic speed prediction. *IEEE Access* 7 (2019), 30050–30057.

[20] Aleksey Ogulenko, Itzhak Benenson, Itzhak Omer, and Barak Alon. 2021. Probabilistic positioning in mobile phone network and its consequences for the privacy of mobility data. *Computers, Environment and Urban Systems* 85 (2021), 101550.

[21] S. Oya, C. Troncoso, and F. Pérez-González. 2019. Rethinking Location Privacy for Unknown Mobility Behaviors. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. 416–431. https://doi.org/10.1109/EuroSP.2019.00038

[22] Anuj S Saxena, Siddharth Dawar, Vikram Goyal, and Debajyoti Bera. 2019. Mining top-k trajectory-patterns from anonymized data. *arXiv preprint arXiv:1912.01861* (2019).

[23] Sina Shaham, Ming Ding, Bo Liu, Shuping Dang, Zihuai Lin, and Jun Li. 2020. Privacy preserving location data publishing: A machine learning approach. *IEEE Transactions on Knowledge and Data Engineering* 33, 9 (2020), 3270–3283.

[24] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[25] Anh Tuan Truong, Quynh Chi Truong, and Tran Khanh Dang. 2010. An adaptive grid-based approach to location privacy preservation. *Advances in intelligent information and database systems* (2010), 133–144.

[26] Jungho Um, Hyeongil Kim, Youngho Choi, and Jaewoo Chang. 2009. A new grid-based cloaking algorithm for privacy protection in location-based services. In *2009 11th IEEE International Conference on High Performance Computing and Communications*. IEEE, 362–368.

[27] Yonghui Xiao and Li Xiong. 2015. Protecting Locations with Differential Privacy under Temporal Correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 1298–1309. https://doi.org/10.1145/2810103.2813640

[28] Toby Xu and Ying Cai. 2009. Feeling-based location privacy protection for location-based services. In *Proceedings of the 16th ACM conference on Computer and communications security*. 348–357.

[29] Dingqi Yang, Daqing Zhang, Bingqing Qu, and Philippe Cudré-Mauroux. 2016. PrivCheck: Privacy-preserving check-in data publishing for personalized location based services. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 545–556.

[30] Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual International Conference on Mobile Computing and Networking*. ACM, 145–156.

[31] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 791–800.