01 Jun 2015

Entropy-Based Privacy against Profiling of User Mobility

Alicia Rodriguez-Carrion

David Rebollo-Monedero

Jordi Forne

Celeste Campo

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/comsci_facwork/589

Recommended Citation

A. Rodriguez-Carrion et al., "Entropy-Based Privacy against Profiling of User Mobility," *Entropy*, vol. 17, no. 6, pp. 3913-3946, MDPI AG, Jun 2015.
The definitive version is available at https://doi.org/10.3390/e17063913

Entropy **2015**, *17*, 3913-3946; doi:10.3390/e17063913

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Entropy-Based Privacy against Profiling of User Mobility

Alicia Rodriguez-Carrion [1,*], David Rebollo-Monedero [2], Jordi Forné [2], Celeste Campo [1], Carlos Garcia-Rubio [1], Javier Parra-Arnau [2] and Sajal K. Das [3]

[1] Department of Telematic Engineering, University Carlos III of Madrid, Avda. Universidad 30, E-28911 Leganés, Madrid, Spain; E-Mails: celeste@it.uc3m.es (C.C.); cgr@it.uc3m.es (C.G.-R.)

[2] Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), Campus Nord, C. Jordi Girona 1-3, 08034 Barcelona, Spain; E-Mails: david.rebollo@entel.upc.edu (D.R.-M.); jforne@entel.upc.edu (J.F.); javier.parra@entel.upc.edu (J.P.-A.)

[3] Computer Science Department, Missouri University of Science and Technology, 325B Computer Science Building, Rolla, MO 65409-0350, USA; E-Mail: sdas@mst.edu

* Author to whom correspondence should be addressed; E-Mail: arcarrio@it.uc3m.es; Tel.: +34-91-624-6234.

Academic Editor: Kevin H. Knuth

Received: 9 March 2015 / Accepted: 8 June 2015 / Published: 10 June 2015

Abstract: Location-based services (LBSs) flood mobile phones nowadays, but their use poses an evident privacy risk. The locations accompanying the LBS queries can be exploited by the LBS provider to build the user profile of visited locations, which might disclose sensitive data, such as work or home locations. The classic concept of entropy is widely used to evaluate privacy in these scenarios, where the information is represented as a sequence of independent samples of categorized data. However, since the LBS queries might be sent very frequently, location profiles can be improved by adding temporal dependencies, thus becoming mobility profiles, where location samples are not independent anymore and might disclose the user's mobility patterns. Since the time dimension is factored in, the classic entropy concept falls short of evaluating the real privacy level, which depends also on the time component. Therefore, we propose to extend the entropy-based privacy metric to the use of the entropy rate to evaluate mobility profiles. Then, two perturbative mechanisms are considered to preserve locations and mobility profiles under gradual utility constraints. We further use the proposed privacy metric and compare it to classic ones to evaluate both synthetic and real mobility profiles when the perturbative methods proposed are applied. The results prove the usefulness of the proposed metric for mobility profiles and the need for tailoring the perturbative methods to the features of mobility profiles in order to improve privacy without completely loosing utility.

Keywords: location-based services (LBSs); entropy; privacy; perturbative methods; location history

1. Introduction

Recent years have witnessed the growth of a rich variety of information and communication technologies. As a result, users enjoy applications striving to tailor information exchange functionality to their specific interests. Examples of these applications comprise location-based services (LBSs), personalized web search and multimedia recommendation systems. In the specific example of LBSs, they open up enormous business opportunities, encompassing intelligent personal assistants, concierge and emergency services, entertainment and advertising, among others. These services are also the result of recent advances in positioning techniques, such as the global positioning system (GPS), location-based social networks (where users tag significant places) or location techniques based on cellular or WiFi networks (where users' locations are inferred by associating the signal strength received by their mobile phone with information about the closest transmission points and their position).

Many of these location services build upon, or lend themselves to, the creation of user profiles. However, user profiles by themselves, but especially when combined across several information services, pose evident privacy and security risks. On the other hand, it is precisely the availability to a system of such sensitive information that enables such intelligent functionality. Therefore, the need for preserving privacy without compromising the utility of the information emerges naturally. Hard privacy [1] is one of the existing privacy-enhancing technologies (PETs) [2] that consists of the preservation of the privacy by the user itself by minimizing, obfuscating or perturbing the information released, without the requirement of trusted intermediaries. In principle, by perturbing the confidential data prior to its disclosure, users attain a certain degree of privacy, at the expense of degrading the system performance (or utility). The existence of this inherent compromise is a strong motivation to develop quantifiable metrics of privacy and utility and to design practical privacy-enhancing, data-perturbative mechanisms achieving serviceable points of operation in this privacy-utility trade-off.

This work focuses on the specific scenario of the privacy risk associated with the profiling of user mobility arising from the use of LBSs and user locations based on the identifier of the cell to which the user's phone is attached to (as opposed to the numerous privacy analysis on GPS-based location). When using LBSs in mobile phones, such as weather, traffic or news widgets, to name a few of the most basic services that depend on the user location, the user's phone sends, quite frequently, a service request together with the user location, aiming to obtain the most up-to-date information. For this kind of service, it is sufficient to know a coarse precision location; thus, cell-based location is suitable whilst diminishing the battery consumption with respect to the GPS option. The LBS provider may, then, collect or disclose to third parties sensible data related to the locations visited by the user. In this work, we distinguish two types of profiles that can be built from the collection of locations sent to the LBS provider. We define the first one as the location profile, and it consists of the set of locations visited by the user and the visit frequency for each one. This profile may disclose implicit information related to the user: her home and work locations; if she has children (the number of visits to a kindergarten or school is high); if she may suffer from some chronic disease (the number of visits to a hospital is high); if she travels much (there are visits to locations located in many different countries); among others. In these cases, an attacker aims at obtaining the most accurate probability distribution of the visits to each location. Then, it would be easier for the attacker to derive the implicit information enclosed in the location profile if a few locations concentrate many more visits, *i.e.*, if the location profile is as different as possible than a uniform distribution. There exist several metrics to measure privacy in this type of scenario, where a set of labeled data exposes the user profile. Some of them are based on the concept of entropy of a set of independent samples, but to the best of our knowledge, it has never been applied to the specific case of sequences of cell-based locations.

Furthermore, we define a second type of profile that can be built by taking advantage of the frequent LBS requests mobile phones usually send to obtain the updated information related to their location. We denote it as the mobility profile, and it is defined as the temporal sequence of locations visited by the user. Therefore, the stress on this profile lies on the correlations among the visited locations, instead of considering the locations as independent events. In this case, an attacker will aim at correctly predicting the next location of the user, given her past history of locations. With this profile, the adversary could derive more refined information due to the knowledge of temporal dependencies. An innocent example of personal mobility information disclosure might be the following one. If the untrusted LBS provider knows, by inspecting the mobility profile, that the user goes from home (first most visited location) to work (second most visited location) and then to a third location near a supermarket, the provider might infer that the user regularly buys products at that supermarket. Therefore, the LBS provider might leak this data to other related services, which can start sending advertisements or offers of different shops offering the same products right before the user goes to her usual supermarket. This behavior, which might result in being very effective for advertisement, is only possible when adding the temporal component to the location profile to transform it into a mobility profile. The problem arising in this situation is that not only the set of visited locations and their visit frequency are the target of a privacy attack, but also the correlations among the visits to those locations constitute a privacy threat. As demonstrated in [3], the correlations among location samples enclose a great deal of information when aiming at predicting the next location of a user. Since this is the target of an adversary, we need to measure privacy taking into account such correlations. However, the classical concept of entropy used for the location profiles does not work on memory processes, because it is only applicable for sequences of independent samples. Therefore, applying privacy metrics based on entropy to a mobility profile does not reflect the real privacy level, since the temporal correlations among locations visits, which represent the main component of a mobility profile, remain ignored. To the best of our knowledge, there exist no privacy metrics addressing the extension from location profiles to mobility profiles. For this reason, we propose to compare the use of classical entropy with respect to an extension of this concept for processes with memory: the entropy rate. This general goal leads to the contributions stated next.

Contribution and Organization

The main contributions of this work are the following ones:

- Jaynes' rationale on maximum entropy methods [4,5] enables us [6] to measure the privacy of confidential data modeled by a probability distribution by means of its Shannon entropy. In particular, one may measure the anonymity of a user's behavioral profile as the entropy of a relative histogram of online activity along predefined categories of interest. Inspired by this principle, we propose the use of Shannon's entropy to measure the privacy of a sequence of places of interest visited by a user (which we will refer to as the user's locations profile), with the caveat that this may only be appropriate for a series of statistically-independent, identically-distributed outcomes.

- Taking this a step further, we tackle the case in which a sequence of location data is more adequately modeled as a stochastic process with memory, representing the (entire or recent) history of a user moving across predefined, discretized locations. We propose extending the more traditional measure of privacy by means of Shannon's entropy, to the more general information-theoretic quantity known as the entropy rate, which quantifies the amount of uncertainty contained in a process with memory. In other words, we put forth the notion of entropy rates as the natural extension of Jaynes' rationale from independent outcomes to time series. Concordantly, we propose the entropy rate as a novel, more adequate measurement of privacy of what we will call user mobility profiles: profiles capturing sequential behavior in which current activity is statistically dependent on the past, as is commonly the case for location data.

- The extension from location to mobility profiles requires a reconsideration of the privacy preserving mechanisms. We propose two simple perturbative methods, previously used for web search applications, looking for their suitability in these two profiling scenarios.

- Finally, we compare the results of calculating the privacy metrics proposed for different theoretical processes of increasing memory, to finally analyze a real location and mobility profile, made up of cell-based locations, which shows the usefulness of the proposed privacy metric. The work ends up with a discussion of different aspects impacting the privacy level obtained and further considerations to improve it in mobility profiling scenarios.

The remainder of the paper is organized as follows. Section 2 presents a detailed study of the state-of-the-art in the two main topics covered in this work: hard-privacy, data-perturbative technologies; and privacy metrics for data perturbation against user profiling. Section 3 states the problem, taking care of the application scenario, the specific privacy model and metrics considered, as well as the perturbative methods to be used. Section 4 exposes the formal analysis of the problem at hand. In Section 5, the experimental data and results are described, together with a discussion on the findings and limitations found. Finally, Section 6 gathers the main conclusions along with some future lines, and the Appendices A to C include the proofs to the mathematical expressions derived in Section 4.

2. Related Work

As exposed in [7], the evolution of LBSs and the associated location techniques lead to a privacy degradation. Anonymous location traces can be identified by correlation with publicly-available databases, thus increasing the possibility of disclosing sensitive data, such as home and work locations [8] or specific points of interest of the user [9]. Therefore, users are exposed to different kinds of attacks (e.g., tracking, localization or meeting attacks, among others [10]) with the available information collected by LBS providers, which can disclose a great deal of the mobility profile of the user. For this reason, privacy enhancement is key in order to tackle the increasing new threats that arise from the evolution of LBSs.

The following is a brief overview of the state-of-the-art of user mobility profiles, since it is the target to protect, along with a review on privacy-enhancing technologies and privacy metrics related to LBSs and profiling of user mobility.

2.1. User Mobility Profiles

Human mobility has been extensively studied, both at a physical level (*i.e.*, the length, pause times and other features of people displacements), as well as at the specific domain level, as the impact of user mobility in the performance of mobile networks and their applications, where our work is enclosed.

There are plenty of works on user profiling, depending on the technology used to collect the mobility traces from which the profiles are made up (e.g., WiFi [11], GSM [12]) and also on the aspects to be reflected in the profile and its immediate application. For instance, in [13], the authors analyze user mobility profiles generated by processing GSM network-based mobility traces and considering the total and daily number of different visited cells, the number of revisits to cells, the frequency of visits or the residence time. Some works, like [14], extract similar features from this same network, like the series of cell identifiers traversed by a user, and use this location history as the mobility profile to detect anomalies in the behavior of the user.

More recent works deal with new types of data [15], such as the location information disclosed in location-based social networks, to create user profiles that can help in content prefetching and rideshare recommendation systems. Going a step further, some works [16] not only deal with profile construction, but also with the comparison metrics among profiles, which can be applied to user recommendations in social networks.

Probably one of the most interesting works on user profiling regarding our work is [17], where the authors incorporate temporal features for the usual spatial profiles, making them more specific to each user, and evaluate the degree to which individuals can be uniquely identified by knowing their spatio-temporal profile. This is directly related to the mobility profiles that we will further define and the increase in the user information they enclose due to the temporal aspects (correlations among locations).

2.2. Privacy-Enhancing Technologies for LBSs

Many different privacy-enhancing techniques focused on LBSs and location profiling can be found in the literature. The statistical disclosure control (SDC) community proposed many of them, aiming to prevent the disclosure of the contribution of specific individuals by inspecting published statistical information. k-anonymity [18,19] is one of the proposed techniques. A specific piece of data on a particular group of individuals is said to satisfy the k-anonymity requirement if the origin of any of its components cannot be ascertained, beyond a subgroup of at least k individuals. The concept of k-anonymity is a widely popular privacy criterion, partly due to its mathematical tractability. However, this tractability comes at the cost of important limitations, which have motivated a number of refinements [20–22].

In the context of statistical databases appears also the concept of differential privacy [23–25]. The idea behind this approach is to guarantee that, after adding random noise to a query, if it is executed on two databases that only differ on one individual, the same answer must be generated with similar probabilities in both databases. Differential privacy is used for LBSs when aggregate location data are published. However, our scenario is that of a single user sending requests to an LBS provider, which is a slightly different case. In order to cope with this difference, the concept of geo-indistinguishability has emerged recently [26,27]. It is a variant of differential privacy for the specific case of LBSs based on the principle that, the closer two locations are, the more indistinguishable they should be. In other words, given two close locations, they should generate the same reported location to the LBS provider with similar probabilities.

Other widely-used alternatives, known as user-centric approaches, rely on perturbation of the location information and user collaboration. In this last context, the authors in [28] propose the collaboration of the users to exchange context information among the interested user and another one who already has that piece of data. This way, many interactions with the LBS provider disappear, thus increasing the location privacy by avoiding as many requests (with the user's location attached to it) to the provider as possible. On the other hand, users' interactions pose in some cases additional privacy risks. That is the case of the effect of co-location in social networks, as demonstrated in [29]. In these situations, even if the user does not disclose her location, she might reveal her friendship and current co-location with a user who does disclose her location. The authors then quantify the impact of these co-location data, deriving an inference algorithm.

Regarding the use of location perturbation techniques, we already introduced the concept of hard privacy [1,30], in its fundamental form of data perturbation carried out locally prior to its disclosure (sometimes referred to as obfuscation), without the requirement of any trusted external party, but inducing a compromise between the privacy attained and the degradation of the utility of the data disclosed for the intended purposes of an information service. A wide variety of perturbation methods for LBSs has been proposed [31]. We only briefly touch upon a few recent ones. In [32], locations and the adjacency between them are modeled by means of the vertices and edges of a graph, assumed to be known by users and providers, rather than coordinates in a Cartesian plane or on a spherical surface. Users provide imprecise locations by sending sets of vertices containing the vertex representing the actual user location. Alternatively, [33] proposes sending circular areas of variable centers and radii in lieu of actual coordinates. Finally, we sketch the idea behind [34]. First, users supply a perturbed location, which the LBS provider uses to compose replies sorted by decreasing proximity. The user may stop requesting replies when geometric considerations guarantee that the reply closest to the undisclosed exact location has already been supplied. Besides these approaches, a number of hard-privacy mechanisms relying on data perturbation have been formulated in an application context wider than LBSs, primarily including online search and resource tagging in the semantic web. Indeed, an interesting approach to provide a distorted version of a user's profile of interests consists of query forgery. The underlying principle is to accompany original queries or query keywords with bogus ones, in order to preserve user privacy to a certain extent. The associated cost relates to traffic and processing overhead, but on the other hand, the user does not need to trust the service provider nor the network. Building on this simple principle, several protocols, mainly heuristic, have been proposed and implemented, with various degrees of sophistication [35–37]. A theoretical study of how to optimize the introduction of bogus queries from an information-theoretic perspective, for a fixed constraint on the traffic overhead, appears in [38]. The perturbation of user profiles for privacy preservation may be carried out not only by means of the insertion of bogus activity, but also by suppression [39]. These approaches constitute the basis of the present work.

Finally, going a step further by preserving not only privacy related to locations understood as a set of independent samples, but also the correlations among locations, the most recent works on location privacy, like [40], take into account the sequential correlation between locations, aiming at protecting the present, past and future locations, as well as the transitions between locations. The authors tackle the problem as a Bayesian Stackelberg problem and use the attacker's estimation error as the privacy metric. This problem is similar to our scenario, since the preserving of the privacy of the correlations among locations is our main concern. However, we tackle the problem with a different approach, using the entropy rate definition. On the other hand, whilst we do not propose any location privacy-preserving mechanism (LPPM) (beyond a couple of naive approaches to demonstrate the usefulness of the proposed privacy metric), the authors of the mentioned work also defined a theoretical framework based on the Bayesian Stackelberg approach to preserve location privacy.

2.3. Privacy Metrics for Data Perturbation against User Profiling

Quantifiable measures of performance are essential to the evaluation of privacy-enhancing mechanisms relying on data perturbation, in terms of both the privacy attained and any degradation of utility. In a recent study on privacy metrics [41], it is shown that many of them may be understood from a unifying conceptual perspective that identifies the quantification of privacy with that of the error in the estimation of sensitive data by a privacy adversary, *i.e.*, privacy is construed as an attacker's estimation error.

Of particular significance is the quantity known as Shannon's entropy [42], a measure of the uncertainty of a random event, associated with a probability distribution across the set of possible outcomes.

Some studies [43–48] propose the applicability of the concept of entropy as a measure of privacy, by proposing to measure the degree of anonymity observable by an attacker as the entropy of the probability distribution of possible senders of a given message in an anonymous communication system. More recent works have taken initial steps in relating privacy to information-theoretic quantities [38,49,50].

A mathematically tractable model of the user profile is a histogram of relative frequencies of visited locations, regarded as a probability distribution, on which we may compute information-theoretic quantities, such as Shannon's entropy. Within the focus of this paper, an intuitive justification in favor of entropy maximization is that it boils down to making the perturbed, observed user profile as uniform as possible, thereby hiding a user's particular bias towards certain visited places. A much richer argumentation stems from Jaynes' rationale behind entropy maximization methods [4,5], more generally understood under the perspective of the method of types and large deviation theory [42]. Under Jaynes' rationale on entropy maximization methods, the entropy of an apparent user profile, modeled by a relative frequency histogram, may be regarded as a measure of privacy or, more accurately, anonymity. The leading idea, proposed in [38,51], is that the method of types from information theory establishes an approximate monotonic relationship between the likelihood of a probability mass function (PMF) in a stochastic system and its entropy. Loosely speaking and in our context, the higher the entropy of a profile, the more likely it is, and the more users behave according to it. This is of course in the absence of a probability distribution model for the probability mass functions, viewed abstractly as random variables themselves. Under this interpretation, entropy is a measure of anonymity, not in the sense that the user's identity remains unknown, but only in the sense that the higher likelihood of an apparent profile, believed by an external observer to be the actual profile, makes that profile more common, hopefully helping the user go unnoticed.

3. Entropic Measures of User Privacy, the Adversary Model and Perturbative Mechanisms

In this section, we describe the scenario where the privacy enhancement techniques will be applied, as well as the theoretical foundation of such techniques. First, we describe how mobility is represented in our scenario, highlighting the difference with respect to the GPS-based mobility representation. Next, once we have defined the data to protect and their representation, we need to know which mechanisms to use in order to enhance the user's privacy. However, in order to evaluate the mechanisms, we first need to define how to measure the privacy level attained. This topic will lead to a discussion on how a concept, such as entropy, is a good privacy measure and how to extend it to the domain of time series through the use of entropy rates. Finally, after defining a quantitative measure of privacy, we propose a perturbative method to enhance user's mobility data privacy under certain utility constraints.

3.1. User Mobility Profiling and the Adversary Model

Users of an LBS disclose trajectories, *i.e.*, sequences of positions, to a service provider. With a small loss of generality for the purposes of user profiling on the basis of behavior, we assume that those positions are not treated in the form of space coordinates, but categorized into a predefined, finite set of labeled symbolic locations. The movement scenario is divided into different regions, each one tagged with a unique identifier. The user moves across this scenario, and each time she enters a different region, the identifier corresponding to that region is recorded as what is known as location history or trace, L. This kind of representation allows recording sequences, such as locations represented by GSM/UMTScells, WiFi coverage areas or sequences of concrete places (office, home, market, gym, *etc.*). This assumption will enable us to model trajectories as random processes with samples distributed in a finite alphabet. In Figure 1, we can see the track of a user that corresponds to the location history $L = afebdihgkjnml$.

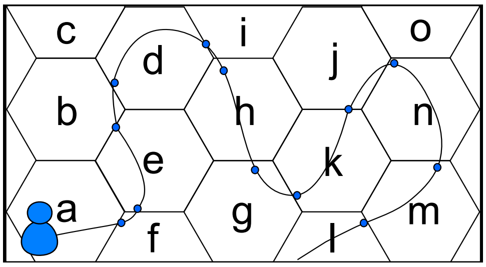

Figure 1. Movement scenario divided in regions and the trajectory followed by the user.

Further, the data contained in a user's trace allow us to define two types of user profile, the location profile and the mobility profile. In the following subsections, we define and comment on these two types of profiles, along with their corresponding adversary models. In the description of those profiles, we describe the connection with the concepts of entropy and the entropy rate as privacy criteria. Additional details on the role of those information-theoretic quantities are provided in Section 3.2.

3.1.1. Location Profile

The location profile is defined as the probability distribution of the visits to each of the locations in the set of visited locations of the user, *i.e.*, the relative frequency of visits of the user's visited location set. This is analogous to the histogram of the relative frequency of the different search categories, in the case of the web search presented in [38,52]. This profile reveals information related to different locations, independently of the rest of the visited locations and correlations among them. For instance, an attacker may be interested in knowing the probability distribution of the visits in order to know several pieces of related data, such as: home or work locations, which are demonstrated to be very easy to derive [3,12], even when the attacker has access to just a few LBS requests [53]; if the user travels to many different countries; if the user usually visits (the relative visit frequency is high) some hospital, religious or political organization, children school, sports center, among others. The attacker, say the LBS provider or a third party to whom the provider relinquishes the user location profile, might use this information to provide personalized advertisement or vary prices depending on the user's demand (e.g., if the frequency of the cumulative visits to locations in a different country to the one with the highest number of visits is high, it can be derived that the user travels frequently, thus she will be prone to book flights at higher prices, because traveling might be part of her work). A high number of visits to a hospital or a religious or political-related venue can have also an impact when looking for jobs or insurances.

- Definition (location profile): Let L be a random variable (r.v.) representing the location of a given user, from an alphabet of predefined location categories $\mathscr{L}$. The time of the location referred to is chosen uniformly at random. We model the location profile of said user as the probability distribution of L; precisely, the probability mass function (PMF) p_L of the discrete r.v. L. Thus, $p_L(l)$ is the probability that the user is at location $l \in \mathscr{L}$ at any given time. In other words, $p_L(l)$ represents the relative frequency with which the user visits this location.

- Example: Examples of location categories that may characterize the behavioral profile of a user include categories, such as "work", "home", "car", "subway", "restaurant", "movie theater" and "out of town". These could be inferred from geographical locations with the help of an appropriate map. Figure 2 depicts a simple example of a location profile on a location alphabet with a few categories.

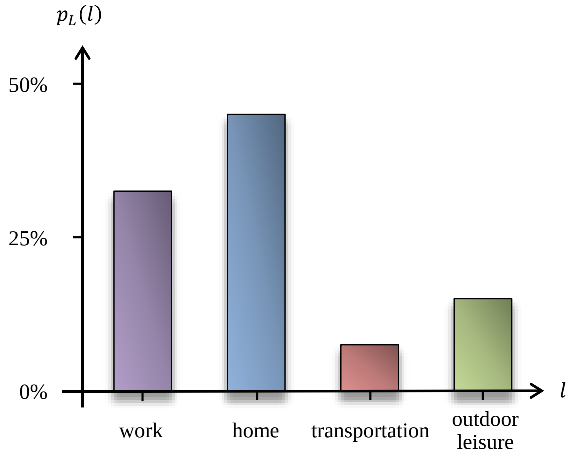

Figure 2. Example of location profile p_L as a probability mass function (PMF) or histogram of relative frequencies on a simple location alphabet $\mathscr{L} = \{$"work", "home", "transportation" and "outdoor leisure"$\}$, inferred from geographical locations.

- Adversary model: The adversary model for the location profile is, in this case, estimating the visit probability distribution as accurately as possible, by inspecting the locations attached to the LBS requests. To this end, the adversary could utilize a maximum likelihood estimate of the distribution, directly as the histogram of relative frequencies, simply by counting observed locations, or any other well-known statistical techniques for the estimation of probability distributions, such as additive or Laplace smoothing.

  Our adversary model contemplates what the attacker is after when estimating those location profiles. According to [6] and in line with the technical literature of profiling [54,55], we assume the attacker's ultimate objective behind profiling is to target users who deviate significantly from the typical location profiles. This is known as individuation, meaning that the adversary aims at discriminating a given user from the whole population of users or, said otherwise, wishes to learn what distinguishes that user from the other users.

  We would like to remark that our interpretation of Shannon's entropy as a measure of profile likelihood is clearly consistent with the assumptions made about the adversary and, in particular, with its objective for constructing location profiles. Specifically, the higher the Shannon entropy of a location profile, the larger the number of users sharing this location pattern and, therefore, the less interesting is the profile to an attacker assumed to target peculiar users. We hasten to stress that the Shannon entropy is, accordingly, a measure of anonymity, rather than privacy, in the sense that the obfuscated information is the uniqueness of the profile behind such location patterns, rather than the actual profile itself.

  Another interpretation of Shannon's entropy as an anonymity metric stems from the intuitive observation that the higher the entropy of the distribution, informally speaking, the flatter the distribution and the less information the attacker could derive about predictable locations. In other words, if all of the locations have the same visit frequency, the attacker can know the visited locations, but not which of them is more important.

3.1.2. Mobility Profile

The mobility profile is defined as the joint probability of visited locations over time or, equivalently, as the sequence of conditional probabilities of the current location, given the past history of locations. In this case, locations are not considered independently as in the user's location profile, but the most important component is the correlation among different locations, *i.e.*, the short- and long-range temporal dependencies among them. In this case, an attacker will aim at predicting the next location that the user will visit, given the past history. The predictions about future locations provide a further refinement for advertisement purposes: the advertiser knows not only which product might be most interesting for the user regarding her visited locations, but also when to offer it for maximizing the impact of the ad. For instance, suggesting some entertainment activity might be more effective if by the mobility profile, the attacker finds that the user did not go from home to work, as usual, which might indicate a weekend or holiday. The adversary's goal is then to be able to predict as accurately as possible the next location of the user, given her past mobility history. There exists many prediction algorithms that can be used to do so [56], and their success depends on the predictability of the mobility history. As demonstrated in [3], the temporal dependencies among the locations visited by the user enclose information that noticeably increases the predictability of the mobility. In that study, the authors define the concept of predictability, closely linked to the entropy rate of the sequence, and demonstrate that it constitutes an upper bound on how much of the time we could correctly predict the next location of the user, given her past mobility history. After analyzing real mobility histories of thousands of users, they conclude that we could correctly predict the next location of a user 93% of the time on average.

- Definition (mobility profile): More precisely, for each user, we define a stochastic process $(L_t)_{t=1,2,\ldots}$ representing the sequence of categorized locations over discretized time instants $t = 1, 2, \ldots$ Concordantly, the corresponding location L_t at time t is a discrete r.v. on the alphabet of predefined location categories $\mathscr{L}$ introduced earlier. We define the mobility profile of the user as the joint probability distribution of locations over time,

  $$p_{L_1\,L_2,\ldots,L_{t-1},\,L_t,\,L_{t+1},\ldots}(l_1\,l_2,\ldots,l_{t-1},\;l_t,\;l_{t+1},\ldots),$$

  which may be equivalently expressed, by the chain rule of probabilities, as the sequence of conditional PMFs of the current location L_t given the past location history $L_{t-1}, L_{t-2}, \ldots$, *i.e.*,

  $$p_{L_t\,|\,L_{t-1},L_{t-2},\ldots}(l_t\,|\,l_{t-1},l_{t-2},\ldots).$$

  Discretized times could be defined in terms of fixed time intervals, such as hours or fractions thereof, or more simply, but less informatively, as times relative to a change in activity of the user, so that the actual logged data are the order of the given locations in time, but not their duration.

- Example: Following up with the simple example of Figure 2, with location alphabet $\mathscr{L} = \{$'work', 'home', 'transportation' and 'outdoor leisure'$\}$, the mobility profile now incorporates time information, in the form of fixed time intervals, say 15 min. In this manner, one could record the average time spent at work, at home, on the road or at various types of outdoor leisure activities and also the mobility patterns involving said locations. With a more detailed location alphabet, one may detect that a user predictably goes to work directly from home, or to the movies, or to a restaurant after work on a given day of the week.

- Adversary model: The mobility profile, characterized by the probability distribution of categorized locations across time, serves to effectively model the knowledge of an adversary about the future locations of a user and raises the concern that motivates our contribution. Since predictability is directly linked to the entropy rate of the mobility profile as a stochastic process, rather than the entropy (the higher the entropy rate, the lower the predictability, as shown in [3]), we could use this information theory concept in order to quantify the privacy of the user mobility profile in such a way that the less predictable a user is (the higher her entropy rate is), the higher her mobility profile privacy will be.

  Analogously to the adversary model for location profiles, we also incorporate here the objective behind such profiling and assume an attacker that strives to find users with atypical mobility profiles. Jaynes' rationale behind the entropy-maximization method allows us to regard the entropy rate (formally presented next) as an anonymity metric that is consistent with this objective.

Table 1 summarizes the main properties of the attacker model considered above, both for location and mobility profiling.

Table 1. Main conceptual highlights of the adversary model assumed in this work.

Who can be the privacy attacker?	Any entity able to profile users based on their location and mobility patterns is taken into account. This includes the location-based service (LBS) provider and any entity capable of eavesdropping on users' location data, e.g., Internet service providers, proxies, users of the same local area network, and so on. Further, we also contemplate any other entity that can collect publicly-available users' data. This might be the case of an attacker crawling the location data that Twitter users attach to their tweets.
How does the attacker model location profiles?	The location profile of a user is modeled as the probability distribution of their locations within an alphabet of predefined location categories. Conceptually, a location profile is a histogram of relative frequencies of user location data across those categories.
How does the attacker represent mobility profiles?	The mobility profile of a user is modeled as the joint probability of visited locations over time. This model is equivalent to the sequence of conditional probabilities of the current location, given the past history of locations.
What is the attacker after when profiling users?	We consider the attacker wishes to individuate users on the basis of their location and mobility patterns. In other words, the adversary aims at finding users who deviate notably from the average and common profile.

3.2. Privacy Model and Additional Discussion on Entropy and the Entropy Rate as Privacy Measures

This work considers an abstract privacy model in which individuals send pieces of confidential data, related to each other in a temporal sequence, to an untrusted recipient. This intended recipient of the data is not fully trusted. In fact, it is regarded as a privacy adversary capable of constructing a profile of sensitive user interests on the basis of the observed activity or prone to leaking such observations to an external party who might carry out the profiling. Disclosure of confidential data to such untrusted recipient poses a privacy risk. However, it is precisely the submission of detailed data on preferences and activity that enables the desired, intelligent functioning of the underlying information system. Although this abstraction is readily applicable to a wide variety of information systems, we focus our exposition on the important example of LBSs.

We have mentioned in our review of the state-of-the-art, Section 2.3, that the anonymity of a profile can be quantified in terms of the Shannon entropy of the probability distribution representing the histogram of relative frequencies profiling user behavior [6]. We also related this interpretation of entropy to our description of privacy attacks based on user profiling in Section 3.1. The work cited argues in favor of the use of this information-theoretic measure capitalizing on Jaynes' rationale for entropy maximization methods. Roughly speaking, Jaynes' argument boils down to postulating that high entropy is more common than low entropy. In the context of privacy, high-entropy profiles are more frequent and, thus, more anonymous.

More sophisticated user profiling may be carried out if the privacy adversary exploits the statistical dependence among location samples over time, in order to infer temporal behavioral patterns. This responds to the observation that the disclosure of a sequence of user locations poses a clear privacy risk, especially when these locations are viewed in conjunction and time is factored in. Examples include answers to questions, such as: Where does a user commonly go after work, before heading back home? On a typical weekend, what is the user's preferred activity after leaving their house? What route does the user usually follow to get to work or back home?

The natural extension of the measurement of anonymity by means of entropy to the case at hand, namely random processes with memory, is the entropy rate, formally defined in Section 4.1. Because the definition of the entropy rate is approximated by the entropy of a large block of consecutive samples (normalized by the number of samples), the very same argument in favor of entropy can be extended to the entropy rate, the latter more suitable to user profiling in terms of trajectory patterns rather than individual locations.

We should remark that entropy has been often proposed as a privacy metric, on the intuitive basis that it constitutes a measure of uncertainty, even though formally speaking, there exist many other such measures, Rènyi entropy or variance, to name a couple. Even though we acknowledge the appropriateness of this intuition, we more formally resort to Jaynes' rationale on maximum entropy methods to argue, as in [6], that more entropic user profiles are also more common and, thus, less idiosyncratic or characteristic of the specific habits of a particular individual. Consequently, the privacy metric proposed here, namely the entropy rate of the stochastic process modeling the sequence of discretized locations, represents a quantifiable measurement of the anonymity of the user, in the sense of commonness, rather than of the confidentiality of the data at hand.

In summary, we argue in favor of measuring the privacy of a user by means of the entropy of its profile, mathematically modeled as a probability distribution across predefined behavioral categories, as a measure of anonymity. We recalled in the state-of-the-art in Section 2.3 and discussed here and in Section 3.1 that Jaynes' rationale on entropy maximization methods enables us to interpret the entropy of a user profile, mathematically modeled as a probability distribution across predefined behavioral categories, as a measure of anonymity. The underlying principle is conceptually simple: more entropic profiles are to be considered *a priori* also more common, so that entropy is a measure of commonness and, thus, anonymity. Taking this a step further, in this work, we propose extending the more traditional measure of privacy by means of Shannon's entropy, to the more general information-theoretic quantity known as the entropy rate, which quantifies the amount of uncertainty contained in a process with memory. In other words, we put forth the notion of entropy rates as the natural extension of Jaynes' rationale from independent outcomes to time series. The corresponding adversary model considers an anonymity attacker striving to ascertain the identity of an individual from its behavioral profile; a harder task when the number of likely candidates abounds.

However, entropy accepts other well-known additional interpretations [42], which we proceed to discuss briefly in the context of privacy. These multiple interpretations make entropy a suitable criterion in diverse privacy applications, each with a corresponding adversary model. Recall, for instance, the well-known interpretation of entropy as a measure of the uncertainty of a random variable and of the entropy rate as a measure of the uncertainty of a random process given its past. We already mentioned that an intuitive justification in favor of entropy maximization is that it boils down to making the perturbed, observed user profile as uniform as possible, thereby hiding a user's particular bias towards certain visited places. Less informally, the fact that entropy is a lower bound on the optimal (Huffman) code length enables us to regard it as a quantifiable measure of the effort of a privacy adversary in obtaining additional bits of information in order to narrow the current uncertainty down to a deterministic outcome, under the perspective of equivalence between source coding and the game of twenty questions. Consistently, Fano's inequality lower bounds the probability of estimation error in terms of a conditional entropy, in the sense of maximum *a posteriori* (MAP) estimation, which can be readily applied to the entropy rate, written as the conditional entropy of a future location of a user given the past history. Here, MAP estimation might be construed as the action taken by a smart privacy attacker.

The arguments above justifies the use of entropy and of the more general information-theoretic quantity known as the entropy rate, as formal, quantitative measures of the effort of a privacy attacker in order to discriminate the identity of an individual from an observed behavior, among others with similar location activity, or in order to characterize and predict its behavior, when the identity is known.

Under the interpretation given by Fano's inequality, for example, the entropy rate $H(L)$, equal to the conditional entropy of the current location L_t given all past locations $L_{t-1}, L_{t-2}, \ldots$, would bound the probability of MAP estimation error of the current location p_e in an alphabet of m possible location categories according to:

$$p_\mathrm{e} \geqslant \frac{H(L) - \log 2}{\log(m-1)}.$$

The higher the entropy rate, the harder to predict the user's current location from her past history. Provided that the underlying stochastic process is stationary, this bound would be the same at any given time.

With regard to estimating identities rather than predicting locations, Jaynes' approximation relates the likelihood $p_L(l)$ of a specific user profile l, represented by a histogram of relative frequencies, based on n sample observations, with the entropy of the l regarded as a distribution, specifically, according to:

$$-\frac{1}{n}\log p_L(l) \simeq D(l\|u) = \log m - H(l),$$

where $D(l\|u)$ denotes the Kullback–Leibler divergence between l and the uniform profile $u = 1/m$ across m location categories. Again, the higher the entropy $H(l)$ of the user profile l, the higher the likelihood $p_L(l)$ of this profile among all users.

The specific unit of information, which would correspond to the basis of the logarithm in the entropic quantity at hand, be it bit (base two), or nat(base e), can be chosen simply by convenience, as it would merely represent a proportionality constant. In Fano's inequality, this constant would be canceled out in the bounding ratio. The binary base, however, would allow the more intuitive interpretation of binary, that is yes/no questions or pieces of information, under the perspective of the aforementioned optimal code length.

The fact that entropy admits multiple interpretations must be regarded as an advantage of a single, unifying theoretical quantity, capable of addressing several privacy models and practical implementation criteria. This multifaceted characterization is not unlike that in other universally-employed formalisms, such as the mean squared error (MSE), which, in addition to mathematical tractability, offer practical interpretations under various standpoints. For instance, on the one hand, the convexity of MSE implies superadditivity, making it a suitable measure of compression quality whenever a few small deviations are preferred to a single large error. On the other hand, MSE may be regarded as a second-order Taylor approximation to any reasonable, symmetric nonlinear measure of distortion when operating at high fidelity.

3.3. Perturbative Mechanisms

Following the reasons stated in the Introduction, particularly motivated by the advantages of hard privacy against the reliance on trusted intermediaries, we shall investigate theoretically and experimentally two data-perturbative strategies prior to the disclosure of trajectories, in order to trade-off usability for privacy. In the first strategy, referred to as uniform replacement from now on, with certain probability, samples are replaced with values drawn according to a uniform distribution over the alphabet of possible categorized locations. In the second mechanism, which will be called improved replacement, the same fraction of samples is replaced, although a more sophisticated policy is employed. Precisely, the replacing samples are drawn from the distribution obtained from the solution to the problem for optimized query forgery developed in [38]. We should point out that because the optimization carried out was originally intended for memoryless processes and anonymity was measured by means of entropy instead of the entropy rate, the aforementioned improved solution need not be optimal whenever the privacy attacker exploits existing statistical dependencies over time. Consequently, both mechanisms we choose to evaluate are merely heuristics.

The probability of replacement is indicative of the degradation in data utility. As we will expose in the next section that the theoretical analysis is equivalent for sample replacement and addition. In this last case, the utility degradation is understood as an increase in the information sent to the LBS provider, thus incrementing the energy consumption of the mobile device and, potentially, the economic cost of data traffic. We consider here applications that can send location samples to the corresponding LBS more frequently than in a normal situation (*i.e.*, where no privacy-enhancing method is applied). That allows one to send fake locations together with the original ones without degrading the service provided, only increasing the cost associated with a more intensive communication. From now on, we will talk about sample replacement, but keeping in mind that it could be extended to sample addition, by slightly changing what we understand by utility in that case. Because sample values may occasionally be replaced by themselves, especially if the number of location categories is small, counting the number of effectively perturbed values is a more adequate measure of utility. While there is ample room for the development of more sophisticated metrics of utility reflecting the quality of the LBS response, the necessarily limited scope of this work prefers to cover the aspects of privacy and perturbation, as the first insightful step towards the problem of privacy-enhanced perturbation of processes with memory.

4. Theoretical Analysis of Perturbative Methods and the Entropy-Based Privacy Metric

4.1. Notation and Information-Theoretic Preliminaries

Throughout the paper, we shall follow the convention of uppercase letters for random variables (r.v.'s) and lowercase letters for particular values they take on. For simplicity, all r.v.'s in this analysis take on values in a finite alphabet. Probability mass functions (PMF) are denoted by p, sub-indexed by the corresponding name of the r.v. when not understood from the context. For instance, we may denote the PMF of an r.v. X at x by $p_X(x)$, or simply by $p(x)$.

We review a few fundamental results from information theory. The reader may refer to [42] for specific details and proofs. The Shannon entropy of an r.v. X with PMF p and finite alphabet $\mathscr{X}$ is written interchangeably as $\mathrm{H}(X)$ or $\mathrm{H}(p)$. Recall that entropy is maximized for the uniform distribution, and for this distribution only, and that the maximum attained is the logarithm of the cardinality of the alphabet. Put mathematically, $\mathrm{H}(p) \leqslant \log|\mathscr{X}|$, with equality if and only if p is the uniform distribution. Throughout this work, all logarithms are taken to base two, and subsequently, the entropy units are bits. Recall also that $\mathrm{H}(p)$ is a strictly concave function of p, in the sense that for any distributions p and p' over the same alphabet and any $\lambda \in [0,1]$,

$$\mathrm{H}((1-\lambda)\,p + \lambda\,p') \geqslant (1-\lambda)\ \mathrm{H}(p) + \lambda\ \mathrm{H}(p'),$$

with equality if and only if $\lambda = 0$, $\lambda = 1$, or $p = p'$.

Let:

$$(X_n)_{n\in\mathbb{Z}} = \ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$$

be a stationary random process with samples defined on a common alphabet $\mathscr{X}$. Stationarity implies that both the entropy sequences $\frac{1}{n}\,\mathrm{H}(X_1,\ldots,X_n)$ and

$$\mathrm{H}(X_n|X_1,\ldots,X_{n-1}) = \mathrm{H}(X_1|X_2,\ldots,X_n)$$

are non-increasing and have a common limit, called the entropy rate, denoted here by $H_R(X)$. For n large, either of these entropy quantities constitutes an arbitrarily accurate approximation to the entropy rate of the process. We can compute these quantities by choosing an appropriate value of n, such that the blocks capture the correlations of the process and calculating $p_X(X_n|X_1,\ldots,X_{n-1})$ as the number of blocks $X_1,\ldots,X_n$ normalized by the total number of blocks of length n, then applying the previous formula $H_R(X) = -\sum p_X(X_n|X_1,\ldots,X_{n-1})\log p_X(X_n|X_1,\ldots,X_{n-1})$.

Stationarity also implies that the samples of the process are identically distributed according to a common PMF. When, in addition, they are statistically independent, the process, or the samples thereof, is then called independent, identically distributed (i.i.d.). More colloquially, a process with independent samples is called memoryless or without memory. For an i.i.d. process, entropy rate and the entropy of individual samples coincide, that is $H_R(X) = H_R(X_n)$. For a general stationary process $H_R(X) \leqslant H_R(X_n)$, with equality if and only if the process is memoryless. The highest entropy rate is attained by processes with independent, uniformly-distributed samples, that is $H_R(X) \leqslant \log|\mathscr{X}|$, with equality if and only if the process is uniformly distributed and memoryless.

4.2. Perturbative Mechanisms

Again, consider a stationary random process $(X_n)_{n\in\mathbb{Z}}$ with samples distributed on a common finite alphabet $\mathscr{X}$. We shall argue elsewhere that entropy rate is an appropriate privacy measure. We propose two alternative privacy-enhancing perturbative mechanisms, in which individual samples of the random process X_n are replaced with X'_n, with probability ρ and independently from each other, as follows.

- Uniform replacement: X'_n is drawn uniformly from $\mathscr{X}$.
- Improved replacement: X'_n is drawn according to the distribution obtained as the solution to the maximum-entropy problem of [38].

Even though [38] was meant for sample addition rather than replacement, the mathematical formulation turns out to be completely equivalent. However, we should hasten to point out that the optimality guarantee of the cited work applies to entropies of individual samples, but not entropy rates in general processes with memory. Consequently, the two alternative mechanisms described above are merely heuristic in the context of this work. In both cases, the resulting perturbed process $(X'_n)_{n\in\mathbb{Z}}$ is clearly stationary.

We shall call ρ the replacement rate. Because sample values may be conceivably replaced with themselves, we propose the following utility measure, which more accurately reflects the actual impact of the perturbative mechanism. Precisely, we define the perturbation rate $\delta = \mathrm{P}\{X_n \neq X'_n\}$, constant with n on account of the stationarity of the processes involved, and observe that $\delta \leqslant \rho$, as only replaced samples may be effectively perturbed, that is actually different.

Even in the heuristic called improved replacement, the samples to be replaced are chosen randomly and replaced independently of their original value. A truly optimal strategy, however, should choose which samples to replace, exploit the statistical model of the memory of the process and be optimized for δ rather than ρ as a measure of utility. The scope of this work is limited to the analysis of the heuristic mechanisms described, as the first step towards shedding some light on the problem of designing perturbative strategies for processes with memory and with a truly optimal privacy-utility trade-off (or privacy-cost qualitatively speaking if we would consider sample addition).

4.2.1. Uniform Replacement

We prove that uniform replacement on stationary processes with a strictly positive replacement rate will always increase the entropy rate, unless the original process is uniformly distributed and memoryless.

Lemma 1. *Let S and U be independent r.v.'s, the latter uniformly distributed on the alphabet of the former. Let T be a third r.v., in general statistically dependent on S. Take $S' = U$ with probability ρ, independently of S and T, and $S' = S$ otherwise. Then, $\mathrm{H}(S'|T) \geqslant \mathrm{H}(S|T)$, with equality if and only if either $\rho = 0$, or else S is uniform and independent of T (refer to Appendix A for the demonstration).*

Theorem 1. *Let $X = (X_n)_{n\in\mathbb{Z}}$ be a stationary random process with samples distributed on a common finite alphabet $\mathscr{X}$. Although the process X itself need not be independent, each of its samples X_n is altered completely independently as follows. Each sample X_n is replaced by another r.v. U_n, uniformly drawn from the alphabet $\mathscr{X}$, with probability ρ, and left intact otherwise. Let $X' = (X'_n)_{n\in\mathbb{Z}}$ be the resulting process, also stationary. Then, for any $m \geqslant 0$,*

$$\mathrm{H}(X'_0|X'_{-1},\ldots,X'_{-m}) \geqslant \mathrm{H}(X_0|X_{-1},\ldots,X_{-m}),$$

with equality if and only if either $\rho = 0$, or else X_0 is uniform and independent of $X_{-1},\ldots,X_{-m}$. The same inequality holds in the limit of $m \to \infty$ yielding entropy rates, that is $\mathrm{H}(X') \geqslant \mathrm{H}(X)$, with equality if and only if either $\rho = 0$, or else X is uniformly distributed and memoryless (refer to Appendix B for the demonstration).

4.2.2. Uniform *versus* Improved Replacement

We show that in the case of memoryless processes that are not originally uniform, improved replacement will require a lower replacement rate to achieve maximum entropy than that demanded by uniform replacement. We shall also see that, when the cardinality of the alphabet is large, the perturbation rate approaches the replacement rate.

In the perturbative mechanisms described earlier, define the critical replacement rate ρ_{crit} to be the replacement rate ρ required for the entropy rate $\mathrm{H}(X')$ of the perturbed process $(X'_n)_{n\in\mathbb{Z}}$ to attain its maximum possible value $\log|\mathscr{X}|$, achievable only when X' becomes memoryless and uniformly distributed. Denote by δ_{crit} the corresponding critical perturbation rate. Write:

$$p_{\text{max}} = \max_{x\in\mathscr{X}} p(x) \geqslant \tfrac{1}{|\mathscr{X}|},$$

with equality if and only if X is uniformly distributed.

Theorem 2. *Assume the nontrivial case in which the original process X is not already independent, uniformly distributed.*

In uniform replacement,

$$\begin{aligned}\delta &= \rho\left(1 - \tfrac{1}{|\mathscr{X}|}\right),\\ \rho_{\text{crit}} &= 1,\\ \delta_{\text{crit}} &= 1 - \tfrac{1}{|\mathscr{X}|}.\end{aligned}$$

In improved replacement, for any $\rho \geqslant 1 - \frac{1}{|\mathscr{X}|p_{\max}}$,

$$\delta = (1-\rho)\sum_x p(x)^2 + \rho - \tfrac{1}{|\mathscr{X}|}.$$

If the original process is i.i.d.,

$$\begin{aligned}\rho_{\text{crit}} &= 1 - \tfrac{1}{|\mathscr{X}|p_{\max}},\\ \delta_{\text{crit}} &= 1 - \tfrac{1}{|\mathscr{X}|} - \tfrac{1}{|\mathscr{X}|p_{\max}}\left(1 - \sum_x p(x)^2\right).\end{aligned}$$

Otherwise, in the general case of processes with memory,

$$\rho_{\text{crit}} = 1 \text{ and } \delta_{\text{crit}} = 1 - \tfrac{1}{|\mathscr{X}|}.$$

(refer to Appendix C for the demonstration).

Recall [42] that the Rényi entropy of order α of a discrete r.v. X with PMF p_X is defined as:

$$\mathrm{H}_\alpha(X) = \frac{1}{1-\alpha}\log\sum_x p_X(x)^\alpha. \tag{1}$$

The value $\sum_x p(x)^2 = \mathrm{E}\, p(X)$ in the theorem is directly related to the Rényi entropy of order two of $p(x)$, called the collision entropy. The sum of squared probabilities above is minimized for the uniform distribution and maximized for a degenerate distribution, where the associated r.v. takes on a single value with probability one.

Observe that in the case of uniform replacement, a large alphabet $|\mathscr{X}|$ implies that the perturbation rate will approach the replacement rate, that is $\delta \simeq \rho$, because of the unlikelihood of replacing a sample by itself. In the case of improved replacement, the approximation requires not only $|\mathscr{X}| \gg 1$, but also $\sum_x p(x)^2 \ll 1$ and only holds for sufficiently large ρ.

4.3. Entropy Estimation

As previously shown, entropy could be a suitable privacy metric, but we should pay attention to the estimator used. Depending on the concrete application or data to focus on, the entropy estimation might be different. In the case of human mobility, the location traces (that lead to locations and mobility profiles) have specific features to take into account when estimating their entropy: strong long-range time-space dependencies, high probabilities of returning to some highly-frequented locations [12], the high number of different visited places (the cardinality of the alphabet), among others.

Bearing these features in mind, we could come up with different entropy estimates, as described in [3], each one of them accounting for different dependencies. As we shall see next, two of these estimates will be the Hartley entropy and the Shannon entropy. Throughout this subsection, these entropies will be denoted by $\mathrm{H}_0(X)$ and $\mathrm{H}_1(X)$ to emphasize their connection with the Rényi entropy, a family of functionals widely used in information theory as a measure of uncertainty. Particularly, from (1), it is straightforward to see that, when $\alpha = 0$, Rényi's entropy boils down to Hartley's. In the limit when α approaches one, this family of functionals reduces to Shannon's entropy.

- Hartley entropy, $\mathrm{H}_0(X)$, is the maximum attainable entropy value. We should recall that entropy is maximized for the uniform distribution and for this distribution only and that the maximum attained is the logarithm of the cardinality of the alphabet. Put mathematically:

  $$\mathrm{H}_0(X) \leqslant \log |X| \tag{2}$$

  with equality if and only if X is drawn from the uniform distribution. Applied to our case, it would be calculated considering the probability mass function of the locations trace (since no temporal dependencies are considered) to be a uniform distribution of $\mathscr{X}$ different symbols (locations). This entropy represents the highest possible uncertainty, as it does not take into account temporal aspects nor the number of visits accumulated by each location.

- Shannon entropy, $\mathrm{H}_1(X)$, is calculated as:

  $$\mathrm{H}_1(X) = -\sum_i p_X(x_i) \log p_X(x_i) \tag{3}$$

  In our scenario, $p_X(x_i)$ is the probability of visiting location x_i, which can be computed as $p_X(x_i) = \frac{N_{x_i}}{N}$, where N_{x_i} is the number of visits received by location x_i, and N is the total number of visits (*i.e.*, the length of the movement history). Shannon entropy considers the correlations in the location visits frequencies, thus being lower (or equal if the probability to visit each location is the same) than $\mathrm{H}_0(X)$. Actually, this entropy would be lower than $\mathrm{H}_0(X)$, such a PMF being less uniform (*i.e.*, as some locations receive many more visits than other ones). Location profiles (because no temporal dependencies are considered) behave precisely like this: some locations corresponding to home or work unite the majority of visits, whilst the rest of the locations are much less visited.

- Entropy rate, $H_R(X)$, comes to the scene when dealing with stationary processes, as pointed out in Section 4.1. It takes into account temporal dependencies between samples of the mobility profile (in this case, we consider the mobility instead of the locations profile, because the time dependencies must be considered). Since $H_R(X)$ takes into account more correlations of the profile, the resulting value is lower than the previous ones (there is less uncertainty regarding the next symbol of the profile) or equal if there are no temporal dependencies.

  Applied to our case, we have a finite number of samples of the profile. Therefore, in order to obtain a good estimate of $H_R(X)$, we should choose the optimal block size, n. This block size should be large enough so that the blocks include important long-term temporal dependencies among location samples. However, since the length of the process (*i.e.*, mobility profile) is limited and the cardinality of the alphabet (*i.e.*, the number of different locations) is high, there are not many samples of long blocks. Thus, choosing a block size too large leads to a poor estimate of $p_X(X_1, \ldots, X_n)$. In order to use an appropriate value of n, we could use a well-known entropy rate estimator based on Lempel–Ziv compression algorithms [3,57]. This way, the estimate of the entropy rate can be calculated as:

$$H'_R = \frac{\ln N}{N \sum_i \Delta_i} \tag{4}$$

where Δ_i is the shortest substring starting at position i, which has not been seen before from Position 1 to $i - 1$, N being the number of samples of the profile.

5. Experimental Study: Results and Discussion

In the previous section, we formulated the theoretical problem of privacy-enhancing in processes with and without memory and how we tackle it. In [38], the authors show some results when the mechanisms proposed are applied to web queries, memoryless process and using a small number of categories. In this section, we will see what happens when the scenario switches to LBSs, where the number of categories increases, the probability model underneath becomes more complex and time starts playing an essential role. First, we will show the privacy gain obtained after applying the privacy-enhancing mechanisms to different processes, both synthetic and real, and finally, we discuss the differences that using real location data bring to the generic problem.

5.1. Results

This section collects the results drawn from applying the perturbative mechanisms described in the previous section to two different datasets. On the one hand, we will use several symbol sequences generated from Markov processes and basic alphabets of two symbols. With these data, we will check the performance of the perturbative methods in simple ideal conditions and observe the influence of an increase of the process memory. On the other hand, real traces, taken from the Reality Mining dataset [58], will be processed and compared with the results of Markov processes, since the real scenario can be considered as an extrapolation of simple Markov processes in terms of memory and the cardinality of the alphabet. More precisely, the location history considered collects the sequence of locations visited by a user (each location represented as the identifier of the cell that the user's phone was attached to at each instant) during an academic year. It gathers more than 500 different cells (symbols) and more than 10,000 cell changes (profile samples or location history length).

In order to show the privacy enhancement evolution, each process is perturbed from 0% of replaced samples (*i.e.*, the original symbol sequence) to 100% of replacements (all samples are replaced), as explained in [38]. For each process and percentage of replacements, 10 realizations are averaged. As a general rule, when $\rho = 0$, we have the original process and, therefore, the original (and minimum) entropy value. As ρ increases, the process starts to become a uniform distribution, which is reached when ρ is maximum, *i.e.*, when all samples are replaced by another one using the perturbative methods previously described, and therefore, for the maximum value of ρ, the entropy value should be equal to H_0. We should recall that ρ is the percentage of replacements, but since, sometimes, the replaced sample is equal to the original one, the real replacement rate is δ.

First, we will analyze how different entropy estimates work when applied to different kinds of processes. More precisely, we will study the influence of an increase of the process memory in the entropy estimation, as well as what happens when the process gets away from a uniform distribution, both in terms of Shannon entropy and the entropy rate. For this last case, we will compare two approaches: the estimation by means of block entropies and the Lempel–Ziv-based estimate.

Figures 3–6 show four different processes of 10,000 samples in each of the plots:

- An almost uniform distribution, drawn from an order-one Markov process with $p(1|0) = 0.45$, $p(0|1) = 0.55$, $p(1) = 0.55$, $\mathrm{H}_0 = \mathrm{H}_1 = H_R = 0.993$. This is the base case.

- An i.i.d. (not uniform) process, drawn from an order-one Markov process with $p(1|0) = 0.8$, $p(0|1) = 0.2$, $p(1) = 0.8$, $\mathrm{H}_0 = 1$, $\mathrm{H}_1 = H_R = 0.772$. Here, we keep the process memoryless and change the probability distribution, such that there is a bias towards one of the two symbols of the alphabet.

- A Markov process with $p(1|0) = 0.2$, $p(0|1) = 0.05$, $p(1) = 0.8$, $\mathrm{H}_0 = 1$, $\mathrm{H}_1 = 0.772$, $H_R = 0.374$. In this case, we increase the memory of the process, keeping the cardinality and probability distribution with respect to the second case.

- A real mobility trace taken from the dataset provided by the Reality Mining Project [58]. We can only theoretically know $\mathrm{H}_0 = 8.765$ (drawn from the cardinality of the alphabet, *i.e.*, the number of different symbols representing the locations visited by the user), since the underlying probability distribution is unknown. This means an increase both in the cardinality and the memory of the process, due to the long-range dependencies of human mobility.

For each figure (process), the entropy value evolution with respect to the replacement rates is plotted. The samples are replaced using the uniform perturbative method, *i.e.*, choosing the new sample from the original alphabet of the sequence with the symbols uniformly distributed. Each process has been generated 10 times, and the results shown here are the average value of the entropy calculated in each repetition.

In the first case shown in Figure 3, the process without replacements is already uniform; therefore, there is no evolution in any of the entropy estimates. When the process is not uniform, but still i.i.d., such as the one in Figure 4, H_1 and H_R coincide, as there is no temporal information that can be captured by H_R to lower the uncertainty, but they are lower than H_0 and increase as the replacements turn the process into a uniform one.

Figure 5 shows what happens when the process is not i.i.d. anymore. In this case, H_R is lower than H_1, as it leverages the temporal information present now in the process to lower the uncertainty.

Finally, in Figure 6, we can see what happens when the number of different symbols (locations) increases, as well as the memory of the process. In this case, we have 500 different symbols, which leads to $\frac{500!}{498!} = 249,500$ possible blocks of two symbols to compute H_R using block entropies (the blocks are of two symbols to compare with respect to the Markov processes). As the number of possible blocks is so high and the number of samples is only of 10,000, as the process becomes uniform, more different blocks of two symbols come to the scene. With this number of samples, we do not have even one occurrence of each different block, the probability of which would be $p_X(X_1, \ldots, X_n) = \frac{498!}{500!} = 4*10^{-6}$. Therefore, when computing $H_R(X)$, the values of the elements of the summation are very small, due to the scarcity of occurrences of each possible block. This scarcity becomes more severe as the process tends to uniformity. Thus, $H_R(X)$ decays to near zero as the number of replaced samples increases, as shown in the figure. As we previously explained in Section 4.3, this entropy estimation is biased by the small number of samples available in the location history of the user (even when it comes from a year of location tracking). This is the reason behind considering a different estimator, like the Lempel–Ziv-based one. Figure 6 shows how this estimator obtains more reasonable results. Both $H_R(X)$ and $H'_R(X)$ are equal for the original sequence ($\rho = 0$). However, in order to analyze the privacy improvement, we need an estimate that works well for all of the replacement rate span. We can also observe that, as the cardinality of the alphabet is much higher, it is more difficult to choose the same sample as the original one in each replacement, and therefore, δ is not bounded to 0.5 as in Markov processes of two symbols.

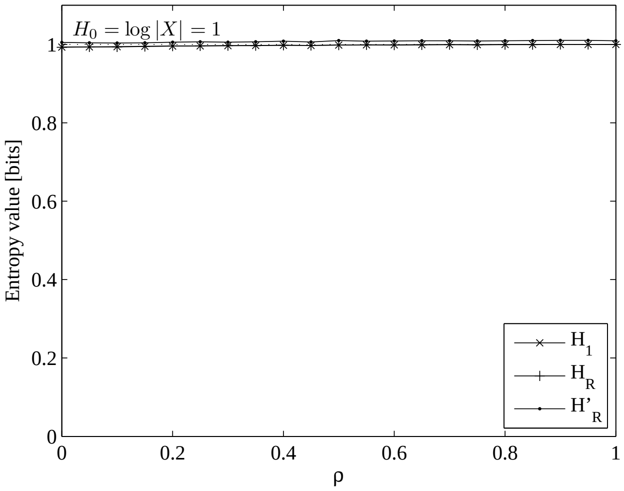

Figure 3. Different entropies for a process drawn from a uniform distribution.

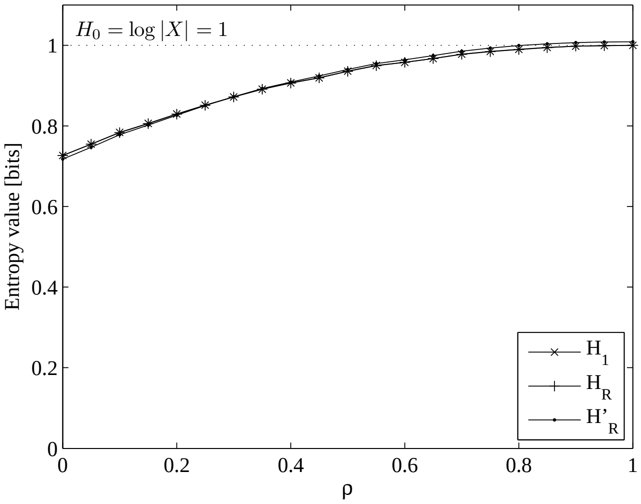

Figure 4. Different entropies for a process drawn from an i.i.d. distribution.

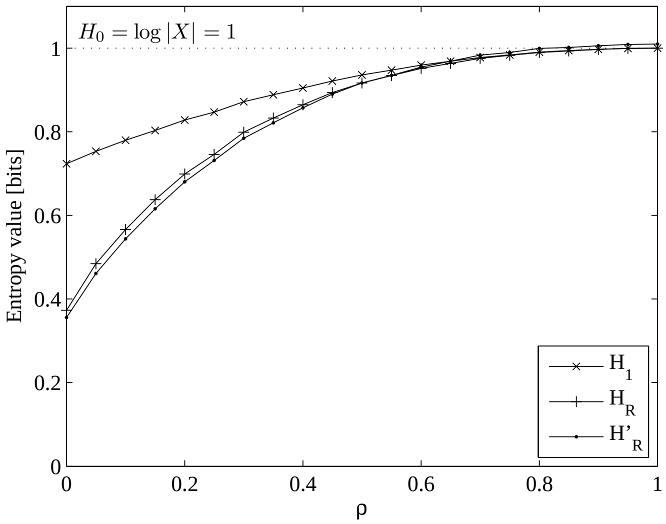

Figure 5. Different entropies for a process drawn from the Markov chain.

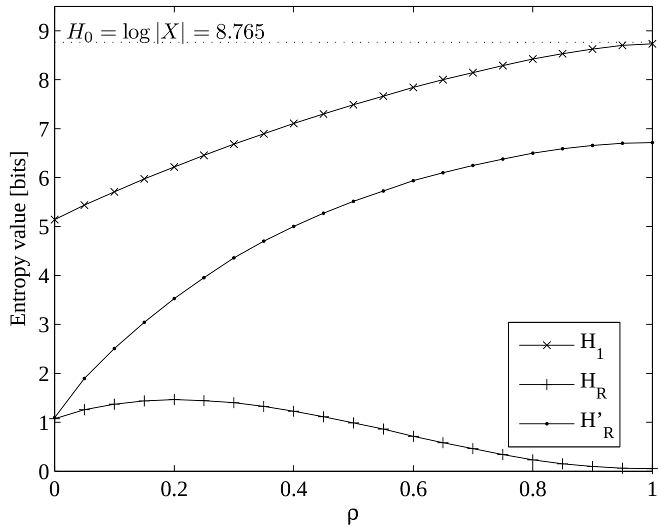

Figure 6. Different entropies for a process drawn from the real mobility trace.

Now that we know how each entropy estimate works for different processes, it is time to apply such estimates to the problem of privacy enhancement.

Figures 7 and 8 represent the privacy level obtained using the perturbative methods described in Section 3.3, for the third Markov process described before (the one with memory) and for the mobility history, respectively. In each figure, four plots can be distinguished: the privacy enhancement in terms of the entropy value, both for Shannon and entropy rate estimates, and for the two perturbative methods considered. Measuring the privacy enhancement by means of two entropy estimates allows one to differentiate the results when only frequency-based information is considered (Shannon's entropy) from the conclusions drawn from time-based data (entropy rate estimate).

For the case of the Markov process with memory in Figure 7, we can see that the privacy enhancement is faster for the improved perturbative method, but only when no time-based information is considered. Besides, it reaches the maximum privacy level (entropy value) when 35% of samples are replaced, a value that lowers up to 25% when the improved perturbative method is used and no temporal information is considered.

When this same analysis is applied to the mobility trace, the results are quite different, as shown in Figure 8. In this case, as the cardinality of the alphabet is so high, it requires 100% of replacements in order to obtain the highest privacy level, when measuring privacy as Shannon's entropy. Besides, the maximum entropy is only achieved when no temporal correlations are considered. In order to get the maximum value for sequence-based data, many more samples would be needed in order to have precise entropy estimation. In this case, it could be checked that the improved perturbation method does not provide faster privacy enhancement for any case.

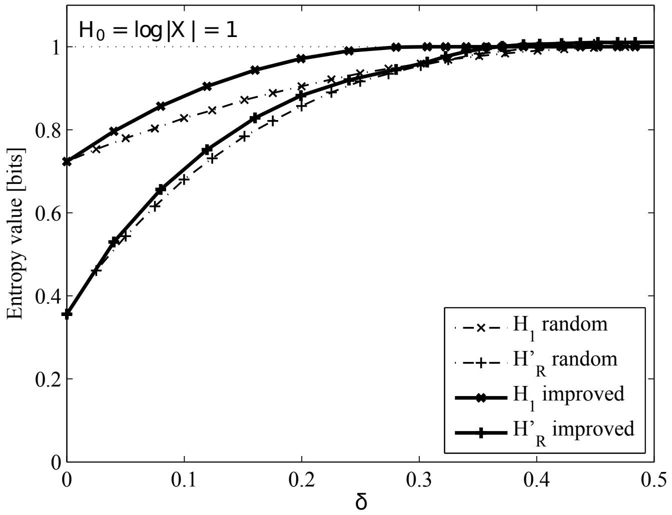

Figure 7. Comparison of perturbative methods for different privacy measures in a Markov process.

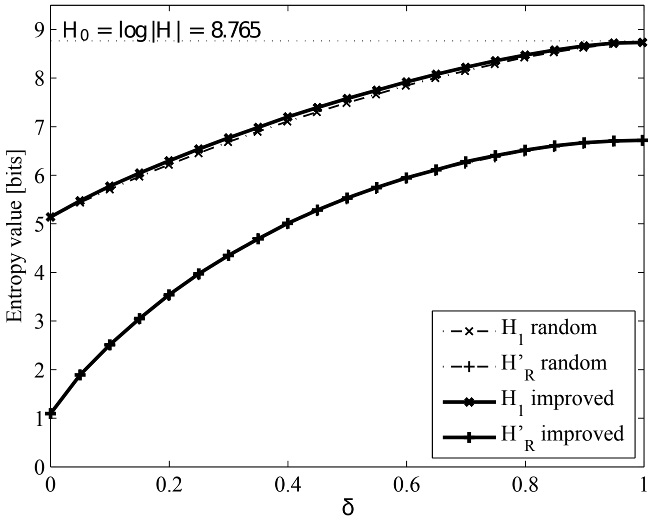

Figure 8. Comparison of perturbative methods for different privacy measures in a real mobility trace.

5.2. Discussion

In the previous figures, the great difference between theory, with simple Markov processes, and real scenarios, such as users mobility profiles, could be observed. However, where do these differences stem from? Although the high cardinality of the alphabet and the complexity of the short- and long-term dependencies of location histories play an important role, the probability distribution underneath the mobility trace is also crucial. A great majority of visits are concentrated in two or three locations, corresponding to home, work and the main points of interest of the user. Therefore, the probability distribution is very biased toward certain locations. The improved perturbative method is based on flattening the underlying distribution with as few replacements as possible in order to get closer to a uniform distribution and, thus, to maximum entropy (*i.e.*, privacy). When the number of categories is not very high and the probability distribution is not very biased to a certain few categories, it is easier to flatten it, as in the case of the Markov processes shown. However, in order to flatten the mobility traces, we would have to compensate the visits to two or three locations through the rest of the 500 different locations visited along the year. Although there are more than 10,000 samples, the cardinality is still very high and would need many more samples to be flattened. This issue is even more critical when considering not only the distribution of the visits, but the sequences of locations. If we consider short-term dependencies (short sequences), we are neglecting important information, and even in this case, the number of combinations is too high to compensate for the number of occurrences of the most repeated sequences. Considering long-term dependencies (long sequences) leads to so many combinations, that there are not enough samples to even calculate a good entropy estimate, even worse if we try to flatten the block probability distribution.

The bias in the visits' probability distribution carries an important consequence: for an attacker, it is very easy to analyze a set of locations and determine where the main points of interest of the user are. Therefore, these are very sensible data that must be masked. The bias can be leveraged in such a way that, instead of trying to flatten all of the distribution, we could focus on the set of the most visited locations and just flatten their number of visits, leaving the least-visited ones as they are. This way, the uncertainty of which of the most visited locations is home or work increases with a smallnumber of replacements. If the number of replacements is not critical or if we could fake the locations to be disclosed, the approach could be to select some of the least visited locations and increase their number of visits to make it comparable to the most visited places. However, as mentioned before, this strategy will require a great number of replacements or additional fake locations. We should remark that adding fake locations incurs battery and data traffic increases, thus being utility-related factors to be taken into account when deciding which perturbative approach to follow.

In the case of location sequences, *i.e.*, focusing on preserving the privacy of the correlations among locations, improving such privacy without compromising the data utility (or avoiding additional cost when adding fake samples instead of replacing the original ones) is more complicated and depends heavily on the application at hand. As we observe in the figures, in order to obtain high privacy levels, the fraction of location samples to change grows fast. Furthermore, the replacements should be done wisely. For example, let a user be walking in Madrid. If, during the user's location sampling, done by the corresponding mobile application communicating with the LBS provider (done every few minutes), the system replaces a location in Madrid by another one in New York City (or just adds the location in New York City in the mobility profile), an attacker could easily detect that it is not possible for a user to make this large jump in such a short period of time. Therefore, this replacement/addition might seem to theoretically improve greatly the privacy level (it is an unexpected movement, thus the entropy rate of the mobility profile would be increased) with little disruption of the utility of the result (because just one location was replaced/added and the system can ignore the result of the associated request, by knowing it is a fake one). However, it would be easy for an attacker to notice the impossibility of the jump, due to the recent past history, and ignore the location in New York. This happens when we are considering a mobility profile, since location profiles on their own just account for the number of visits to each place, leaving unnoticed this kind of impossible large jump between locations in a short period of time. We can devise then a semantics and scale-related problem. What data do we want to preserve? For instance, if only the work/home locations are the ones to be protected, the perturbation methods should focus on replacing or adding samples of the same city repeatedly, so that their frequency is comparable to the one of home/work locations. Since the places could be nearby, it would be more difficult for an attacker to distinguish among the real and fake ones. However, if we want to preserve in which country the user is, the perturbation mechanism needs to be more sophisticated to make the attacker believe that the user might be at any of two countries by creating equally believable location profiles. Since we need to keep the scope of this work bounded, we just analyzed the basic cases and made some considerations about these interesting questions for further research.

As mentioned in Section 2.2, to the best of our knowledge, there is only one other work aiming at preserving the information contained in the transitions among locations [40]. For the sake of completeness, we finish up with a comparison of their results and ours. In Theodorakopoulos *et al.*'s work, the goal is to design the optimal LPPM that takes the real locations to be protected together with the previous fake ones sent to the LBS and selects, probabilistically, the fake locations to send next. In order to obtain the optimal mechanism, they assume that the adversary already knows how the LPPM works and, thus, will respond optimally to it. To design the optimal LPPM under these assumptions, they use a Stackelberg game strategy. The design parameters are the privacy gain function, expressed as the adversary's error in estimating the real user location (*i.e.*, the distance function mapping the value assigned to the difference between the real user location and the one estimated by the adversary) and the quality metric function (*i.e.*, the quality loss incurred when each fake location is sent instead of the corresponding real one). As we can see, the work exposed in [40] focuses on the LPPM design under certain constraints, whereas in our work, the privacy metric is the main focus (since our obfuscation mechanisms are mere heuristics to validate the proposed metric). Their work provides flexibility on the privacy metric to use, as long as it can be expressed in terms of the distance between the location to protect and the estimation of the adversary. Thus, it is dependent on the adversary behavior. In our case, the privacy metric we propose is just dependent on the user mobility and, more specifically, on her predictability. Entropy and the entropy rate are tightly coupled with this predictability concept, as originally stated in [3]. Each user has a different maximum predictability (calculated based on the entropy rate value), which corresponds to the maximum percentage of time a prediction algorithm could correctly predict the next location of the user. Thus, more random users (*i.e.*, users with higher entropy values and, thus, lower predictability) could achieve higher privacy, independently of how well the adversary is able to infer their real locations. In other words, our privacy metric assumes that the adversary applies the best prediction algorithm that could ever exist in order to infer the real location of the user. This is somehow similar to what the authors in [40] assume to design their LPPM: the adversary knows the LPPM and, thus, can respond optimally. In fact, we can see that the results obtained in both cases behave very similarly: the privacy level (in their case, using the hamming distance as an illustrative example) grows as the quality loss increases, until a certain point from which, although increasing the quality (or utility) loss, the achieved privacy remains the same. This point is directly related to the predictability of the user and, as also pointed out in [40], depends on the specific user, and it is not a constraint of the system, but of the user mobility model itself.

Regarding the privacy-utility trade-off, the authors of [40] face it by including a function that reflects the distance among the real location and the fake one that is calculated by their proposed LPPM. Then, the quality (utility) loss is calculated by averaging the result of applying he function to each location that needs to be disclosed for the LBS (and, thus, that can reach the attacker). The distance function reflects which events have more impact on quality loss and also the different quality losses of different outputs (fake locations) for each target location to be protected. However, they do not explicitly consider the semantics of the locations and transitions among them, although they designed their LPPM to face adversaries who can learn how the LPPM works at each step (thus, potentially being able to notice also if the LPPM is replacing samples by other ones far away from the current one). Anyhow, the semantics and correlation-related problem is shown to be crucial when assessing the real privacy obtained by the LPPM and, thus, needs to be further investigated in future works.

6. Conclusions and Future Work

In this work, we have analyzed privacy-enhancing mechanisms based on information theory concepts, such as entropy, applied to locations and mobility profiling scenarios. Starting with synthetic and simple processes, we have shown that the theory applicable to these low alphabet cardinality, memoryless processes cannot be directly applied to more complex cases, such as the mobility profiles of users. Therefore, the remarkable results obtained in the simpler case get degraded until little privacy enhancement is observed, unless utility is completely lost.

The main reasons leading to these results are the increase in the alphabet cardinality (from a few categories to hundreds of visited places by a user) and the temporal dependencies introduced by the fact of considering mobility profiles instead of set of independent samples (location profiles), which leads to the need for using general privacy metrics, such as the one proposed in this work, based on the information theory concept of the entropy rate, in order to consider the temporal dependencies of the mobility profiles. Moreover, the probability density function underneath in the mobility profile of a user is highly biased toward certain frequently-visited places, which makes it difficult to hide these locations just by replacing the rest of the samples by random locations.

As discussed earlier, careful replacement methods should be studied for these special cases. An interesting future research line might be to investigate how to replace samples taking into account the current and past locations, in order to provide reasonable replacements and to exploit the biases toward the most visited locations to flatten the probability distribution, since these locations and their visitation profiles are the keys to identify the user behind such profiles. Another interesting aspect to explore is the usefulness of alternative measures of uncertainty, such as the Rènyi entropy and the variance, in order to assess the privacy of mobility profiles.

Acknowledgments

This work is partially supported by the Spanish Ministry of Science and Innovation through the CONSEQUENCE(TEC2010-20572-C02-01/02) and EMRISCO (TEC2013-47665-C4-4-R) projects. The work of Das was partially supported by NSF Grants IIS-1404673, CNS-1355505, CNS-1404677 and DGE-1433659. Part of the work by Rodriguez-Carrion was conducted while she was visiting the Computer Science Department at Missouri University of Science and Technology in 2013–2014.

Author Contributions

A. Rodriguez-Carrion, C. Campo and C. Garcia-Rubio participated in the conception and development of the main idea, motivation and discussion and contributed mainly in the design of the experiments and manuscript preparation. D. Rebollo-Monedero, J. Fornéand J. Parra-Arnau actively participated in the conception and development of many of the conceptual, theoretical and experimental aspects of the paper, but particularly in the information-theoretic formulation and analysis of the problem investigated. They also made critical revision of the manuscript at all stages of the preparation. S.K. Das thoroughly revised the paper and provided useful feedback for its improvement. All authors have read and approved the final manuscript.

Appendix

A. Proof of Lemma 1

Proof. For each t (with $p(t) > 0$) and each s,

$$p_{S'|T}(s|t) = (1-\rho)\, p_{S|T}(s|t) + \rho\, \tfrac{1}{k},$$

where k is the cardinality of the alphabet of S. Due to the concavity of the entropy and the fact that uniform distributions maximize it, for all t,

$$\mathrm{H}(S'|t) \geqslant (1-\rho)\, \mathrm{H}(S|t) + \rho \log k \geqslant \mathrm{H}(S|t),$$

where $\mathrm{H}(S|t)$ denotes the entropy of S given $T = t$ and similarly for S'. Taking expectations on t, $H(S'|T) \geqslant H(S|T)$. Clearly, equality holds only when $\rho = 0$, or else, when S given t is uniformly distributed, regardless of t, *i.e.*, $p(s|t) = \frac{1}{k} = p(s)$. □

B. Proof of Theorem 1

Proof. We prove the statement for the nontrivial case when $\rho > 0$. In Lemma 1, take $S = X_0$, $S' = X'_0$ and $T = (X_{-1}, \ldots, X_{-m})$; thus

$$\mathrm{H}(X'_0|X_{-1}, \ldots, X_{-m}) \geqslant \mathrm{H}(X_0|X_{-1}, \ldots, X_{-m}),$$

with equality if and only if X_0 is uniform and independent of $(X_{-1}, \dots, X_{-m})$. Next, observe that X'_0 and $(X'_{-1}, \dots, X'_{-m})$ are conditionally independent given $(X_{-1}, \dots, X_{-m})$. Apply the conditional-entropy form of the data processing inequality to write:

$$\mathrm{H}(X'_0|X'_{-1}, \dots, X'_{-m}) \geqslant \mathrm{H}(X'_0|X_{-1}, \dots, X_{-m}),$$

with equality if and only if X'_0 and $(X_{-1}, \dots, X_{-m})$ are conditionally independent given $(X'_{-1}, \dots, X'_{-m})$. Combine both inequalities to immediately conclude the assertions in the theorem regarding m past samples. The claims on the limit of m for entropy rates follow the same proof, with $S = X_0$, $S' = X'_0$ and $T = (X_{-1}, X_{-2}, \dots)$. □

C. Proof of Theorem 2

Proof. In uniform replacement, a sample X_n will be effectively perturbed when replacement occurs, with probability ρ, and the replacement sample U_n does not match the original one. Precisely,

$$\delta = \mathrm{P}\{X_n \neq X'_n\} = \rho(1 - \mathrm{P}\{X_n = U_n\}).$$

Because X_n and U_n are independent and U_n is uniform,

$$\mathrm{P}\{U_n = X_n\} = \mathrm{E}_{X_n} \mathrm{P}\{U_n = X_n | X_n\} = 1/|\mathscr{X}|.$$

If the original process X is not independent, uniformly distributed, all samples must be replaced to make it so, thereby maximizing the entropy rate. Consequently, $\rho_{\text{crit}} = 1$, and δ_{crit} can be obtained from the relationship between ρ and δ above, simply by setting $\rho = 1$.

As for improved replacement, we resort to Theorem 2 in [38] and the concept of critical redundancy, which takes on the value $1 - \frac{1}{|\mathscr{X}| p_{\max}}$ in the notation of this work. According to this, for any $\rho \geqslant 1 - \frac{1}{|\mathscr{X}| p_{\max}}$, the PMF of the replaced samples R_n is:

$$r(x) = \frac{1}{\rho} \frac{1}{|\mathscr{X}|} + \left(1 - \frac{1}{\rho}\right) p(x).$$

Proceeding as in the first part of this proof,

$$\delta = \rho(1 - \mathrm{P}\{X_n = R_n\}),$$

but now:

$$\mathrm{P}\{X_n = R_n\} = \sum_x p(x)\, r(x),$$

from which the expression for δ in the second part of the theorem follows.

For i.i.d. processes, the problem is mathematically equivalent to that formulated in [38], and ρ_{crit} becomes the critical redundancy defined shortly before Theorem 2 in the cited work, in the form expressed in the statement of the theorem we prove here.

The case for processes with memory requires complete replacement to achieve the independence of the samples, not merely uniform distribution, just as in the case of uniform replacement. However, for $\rho = 1$, the replacement strategy R_n becomes uniform, and the analysis for uniform replacement above applies here, as well. □

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Danezis, G. *Introduction to Privacy Technology*; Katholieke University Leuven, COSIC: Leuven, Belgium, 2007.
2. Dingledine, R. *Free Haven's Anonymity Bibliography*; Free Haven Project, Massachusetts Institute of Technology: Massachusetts, MA, USA, 2009.
3. Song, C.; Qu, Z.; Blumm, N.; Barabási, A.L. Limits of Predictability in Human Mobility. *Science* **2010**, *327*, 1018–1021.
4. Jaynes, E.T. On the Rationale of Maximum-Entropy Methods. *Proc. IEEE* **1982**, *70*, 939–952.
5. Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev.* **1957**, *108*, 171–190.
6. Parra-Arnau, J.; Rebollo-Monedero, D.; Forné, J. Measuring the privacy of user profiles in personalized information systems. *Future Gen. Comput. Syst.* **2014**, 53–63.
7. Wicker, S.B. The Loss of Location Privacy in the Cellular Age. *Commun. ACM* **2012**, *55*, 60–68.
8. De Mulder, Y.; Danezis, G.; Batina, L.; Preneel, B. Identification via Location-profiling in GSM Networks. In Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society (WPES '08), Alexandria, VA, USA, 27 October 2008; ACM: New York, NY, USA, 2008; pp. 23–32.
9. Freudiger, J.; Shokri, R.; Hubaux, J.P. Evaluating the Privacy Risk of Location-based Services. In Proceedings of the 15th International Conference on Financial Cryptography and Data Security, Gros Islet, St. Lucia, 28 February–4 March 2011; Springer-Verlag: Berlin, Germany, 2012; pp. 31–46.
10. Shokri, R.; Theodorakopoulos, G.; Le Boudec, J.Y.; Hubaux, J.P. Quantifying Location Privacy. In Proceedings of the 2011 IEEE Symposium on Security and Privacy (SP), Berkeley, CA , USA, 22–25 May 2011; pp. 247–262.
11. Hsu, W.j.; Dutta, D.; Helmy, A. Structural analysis of user association patterns in university campus wireless lans. *IEEE Trans. Mobile Comput.* **2012**, *11*, 1734–1748.
12. Gonzalez, M.C.; Hidalgo, C.A.; Barabasi, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782.
13. Sricharan, M.; Vaidehi, V. User Classification Using Mobility Patterns in Macrocellular Wireless Networks. In Proceedings of the 2011 International Symposium on Ad Hoc and Ubiquitous Computing, 2006 (ISAUHC '06), Surathkal, India, 20–23 December 2006; pp. 132–137.
14. Sun, B.; Yu, F.; Wu, K.; Xiao, Y.; Leung, V. Enhancing security using mobility-based anomaly detection in cellular mobile networks. *IEEE Trans. Veh. Technol.* **2006**, *55*, 1385–1396.
15. Rallapalli, S.; Dong, W.; Lee, G.M.; Chen, Y.C.; Qiu, L. Analysis and applications of smartphone user mobility. In Proceedings of the 2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Turin, Italy, 14–19 April 2013; pp. 235–240.
16. Chen, X.; Pang, J.; Xue, R. Constructing and comparing user mobility profiles. *ACM Trans. Web (TWEB)* **2014**, *8*, 21.

17. Pham, N.; Cao, T. A Spatio-Temporal Profiling Model for Person Identification. In *Knowledge and Systems Engineering*; Springer: Berlin, Germany, 2014; pp. 363–373.
18. Samarati, P.; Sweeney, L. Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression. In Proceedings of the IEEE Symposium on Research in Security and Privacy, Oakland, CA, USA, 3–6 May 1998.
19. Samarati, P. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **2001**, *13*, 1010–1027.
20. Truta, T.M.; Vinay, B. Privacy protection: p-Sensitive k-anonymity property. In Proceedings of the IEEE 22nd International Conference on Data Engineering Workshops (ICDEW), Atlanta, GA, USA, 3–7 April 2006; p. 94.
21. Sun, X.; Wang, H.; Li, J.; Truta, T.M. Enhanced p-sensitive k-anonymity models for privacy preserving data publishing. *Trans. Data Priv.* **2008**, *1*, 53–66.
22. Machanavajjhala, A.; Gehrke, J.; Kiefer, D.; Venkitasubramanian, M. l-Diversity: Privacy beyond k-anonymity. In Proceedings of the IEEE 22nd International Conference on Data Engineering (ICDE), Atlanta, GA, USA, 3–7 April 2006; p. 24.
23. Dwork, C. Differential privacy. In *Encyclopedia of Cryptography and Security*; Springer: Berlin, Germany, 2011; pp. 338–340.
24. Ho, S.S.; Ruan, S. Differential privacy for location pattern mining. In Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, Chicago, IL, USA, 1–4 November 2011; pp. 17–24.
25. Chen, R.; Acs, G.; Castelluccia, C. Differentially private sequential data publication via variable-length n-grams. In Proceedings of the 2012 ACM Conference on Computer and Communications Security, Raleigh, NC, USA, 16–18 October 2012; pp. 638–649.
26. Bordenabe, N.E.; Chatzikokolakis, K.; Palamidessi, C. Optimal geo-indistinguishable mechanisms for location privacy. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, 3–7 November 2014; pp. 251–262.
27. Chatzikokolakis, K.; Palamidessi, C.; Stronati, M. A predictive differentially-private mechanism for mobility traces. In *Privacy Enhancing Technologies*; Springer: Berlin, Germany, 2014; pp. 21–41.
28. Shokri, R.; Theodorakopoulos, G.; Papadimitratos, P.; Kazemi, E.; Hubaux, J.P. Hiding in the Mobile Crowd: LocationPrivacy through Collaboration. *IEEE Trans. Dependable Secur. Comput.* **2014**, *11*, 266–279.
29. Olteanu, A.M.; Huguenin, K.; Shokri, R.; Hubaux, J.P. Quantifying the effect of co-location information on location privacy. Privacy Enhancing Technologies. Springer, 2014, pp. 184–203.
30. Deng, M. Privacy Preserving Content Protection. Ph.D. Thesis, Department of Electrical Engineering, Katholieke Univivresity Leuven, Leuven, Belgium, 2010.
31. Duckham, M.; Mason, K.; Stell, J.; Worboys, M. A formal approach to imperfection in geographic information. *Comput. Environ. Urban Syst.* **2001**, *25*, 89–103.
32. Duckham, M.; Kulit, L. A Formal Model of Obfuscation and Negotiation for Location Privacy. In Proceedings of the 3rd International Conference on Pervasive Computing, Munich, Germany, 8–13 May 2005; Springer-Verlag: Berlin, Germany, 2005; Volume 3468, pp. 152–170.

33. Ardagna, C.A.; Cremonini, M.; Damiani, E.; S. De Capitani di Vimercati.; Samarati, P. Location Privacy Protection Through Obfuscation-Based Techniques. In Proceedings of the Working Conference on Data and Applications Security, Redondo Beach, CA, USA, 8–11 July 2007; Springer-Verlag: Berlin, Germany, 2007; Volume 4602; pp. 47–60.
34. Yiu, M.L.; Jensen, C.S.; Huang, X.; Lu, H. SpaceTwist: Managing the trade-offs among Location Privacy, Query Performance, and Query Accuracy in Mobile Services. In Proceedings of the IEEE 24th International Conference on Data Engineering, 2008, (ICDE 2008), Cancun, Mexico, 7–12 April 2008; pp. 366–375.
35. Kuflik, T.; Shapira, B.; Elovici, Y.; Maschiach, A. Privacy preservation improvement by learning optimal profile generation rate. In *User Modeling*; Springer-Verlag: Berlin, Germany, 2003; Volume 2702, pp.168–177.
36. Elovici, Y.; Glezer, C.; Shapira, B. Enhancing customer privacy while searching for products and services on the World Wide Web. *J. Med. Internet Res.* **2005**, *15*, 378–399.
37. Shapira, B.; Elovici, Y.; Meshiach, A.; Kuflik, T. PRAW—The model for PRivAte Web. *J. Am. Soc. Inf. Sci. Technol.* **2005**, *56*, 159–172.
38. Rebollo-Monedero, D.; Forné, J. Optimal Query Forgery for Private Information Retrieval. *IEEE Trans. Inf. Theory* **2010**, *56*, 4631–4642.
39. Parra-Arnau, J.; Rebollo-Monedero, D.; Forné, J. Optimal Forgery and Suppression of Ratings for Privacy Enhancement in Recommendation Systems. *Entropy* **2014**, *16*, 1586–1631.
40. Theodorakopoulos, G.; Shokri, R.; Troncoso, C.; Hubaux, J.P.; Le Boudec, J.Y. Prolonging the Hide-and-Seek Game: Optimal Trajectory Privacy for Location-Based Services. In Proceedings of the 13th Workshop on Privacy in the Electronic Society, Scottsdale, AZ, USA, 3–7 November 2014; pp. 73–82.
41. Rebollo-Monedero, D.; Parra-Arnau, J.; Diaz, C.; Forné, J. On the Measurement of Privacy as an Attacker's Estimation Error. *Int. J. Inf. Secur.* **2013**, *12*, 129–149.
42. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: New York, NY, USA, 2006.
43. Serjantov, A.; Danezis, G. Towards an Information Theoretic Metric for Anonymity. In Proceedings of the 2nd International Conference on Privacy Enhancing Technologies (PET), San Francisco, CA, USA, 14–15 April 2002; Springer-Verlag: Berlin, Germany, 2002; Volume 2482, pp. 41–53.
44. Díaz, C.; Seys, S.; Claessens, J.; Preneel, B. Towards measuring anonymity. In Proceedings of the 2nd International Conference on Privacy Enhancing Technologies (PET), San Francisco, CA, USA, 14–15 April 2002; Springer-Verlag: Berlin, Germnay, 2002; Volume 2482, pp. 54–68.
45. Díaz, C. Anonymity and Privacy in Electronic Services. Ph.D. Thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2005.
46. Oganian, A.; Domingo Ferrer, J. A posteriori disclosure risk measure for tabular data based on conditional entropy. *SORT* **2003**, *27*, 175–190.
47. Voulodimos, A.S.; Patrikakis, C.Z. Quantifying privacy in terms of entropy for context aware services. *Identity Inf. Soc.* **2009**, *2*, 155–169.

48. Alfalayleh, M.; Brankovic, L. Quantifying Privacy: A Novel Entropy-Based Measure of Disclosure Risk. **2014**, arXiv:1409.2112.
49. Rebollo-Monedero, D.; Forné, J.; Domingo-Ferrer, J. From t-Closeness-Like Privacy to Postrandomization via Information Theory. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1623–1636.
50. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDE), Istanbul, Turkey, 15–20 April 2007; pp. 106–115.
51. Parra-Arnau, J.; Rebollo-Monedero, D.; Forné, J. A Privacy-Preserving Architecture for the Semantic Web based on Tag Suppression. In Proceedings of the 7th International Conference on Trust, Privacy, Security in Digital Bussiness (TRUSTBUS), Bilbao, Spain, 30 August–3 September 2010; pp. 58–68.
52. Parra-Arnau, J.; Rebollo-Monedero, D.; Forné, J.; Muñoz, J.L.; Esparza, O. Optimal tag suppression for privacy protection in the semantic Web. *Data Knowl. Eng.* **2012**, *81–82*, 46–66.
53. De Montjoye, Y.A.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Rep.* **2013**, *3*. doi:10.1038/srep01376.
54. Hildebrandt, M.; Backhouse, J.; Andronikou, V.; Benoist, E.; Canhoto, A.; Diaz, C.; Gasson, M.; Geradts, Z.; Meints, M.; Nabeth, T.; Bendegem, J.P.V.; der Hof, S.V.; Vedder, A.; Yannopoulos, A. *Descriptive Analysis and Inventory of Profiling Practices—Deliverable 7.2*; Technical report, 2005; Available online: http://www.fidis.net/resources/fidis-deliverables/profiling/int-d72000/doc/2/ (accessed on 10 June 2015).
55. Hildebrandt, M., Gutwirth, S., Eds. *Profiling the European Citizen: Cross-Disciplinary Perspectives*; Springer-Verlag: Berlin, Germany, 2008.
56. Rodriguez-Carrion, A.; Garcia-Rubio, C.; Campo, C.; Cortés-Martín, A.; Garcia-Lozano, E.; Noriega-Vivas, P. Study of LZ-Based Location Prediction and Its Application to Transportation Recommender Systems. *Sensors* **2012**, *12*, 7496–7517.
57. Schurmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos Interdiscip. J. Nonlinear Sci.* **1996**, *6*, 414–427.
58. Eagle, N.; Pentland, A.; Lazer, D. Inferring Social Network Structure using Mobile Phone Data. *PNAS* **2009**, *106*, 15274–15278.