Privacy Protection Against User Profiling Through Optimal Data Generalization

César Gil
Department of Telematics Engineering
Universitat Politècnica de Catalunya
Barcelona, Spain
cesar.gil@upc.edu

Javier Parra-Arnau
Department of Telematics Engineering
Universitat Politècnica de Catalunya
Barcelona, Spain
javier.parra@upc.edu

Jordi Forné
Department of Telematics Engineering
Universitat Politècnica de Catalunya
Barcelona, Spain
jordi.forne@upc.edu

***Abstract*—Personalized information systems are information-filtering systems that endeavor to tailor information-exchange functionality to the specific interests of their users. The ability of these systems to profile users based on their search queries at Google, disclosed locations at Twitter or rated movies at Netflix, is on the one hand what enables such intelligent functionality, but on the other, the source of serious privacy concerns. Leveraging on the principle of data minimization, we propose a data-generalization mechanism that aims to protect users' privacy against non-fully trusted personalized information systems. In our approach, a user may like to disclose personal data to such systems when they feel comfortable. But when they do not, they may wish to replace specific and sensitive data with more general and thus less sensitive data, before sharing this information with the personalized system in question. Generalization therefore may protect user privacy to a certain extent, but clearly at the cost of some information loss. In this work, we model mathematically an optimized version of this mechanism and investigate theoretically some key properties of the privacy-utility trade-off posed by this mechanism. Experimental results on two real-world datasets demonstrate how our approach may contribute to privacy protection and show it can outperform state-of-the-art perturbation techniques like data forgery and suppression by providing higher utility for a same privacy level. On a practical level, the implications of our work are diverse in the field of personalized online services. We emphasize that our mechanism allows each user individually to take charge of their own privacy, without the need to go to third parties or share resources with other users. And on the other hand, it provides privacy designers/engineers with a new data-perturbative mechanism with which to evaluate their systems in the presence of data that is likely to be generalizable according to a certain hierarchy, highlighting spatial generalization, with practical application in popular location based services.**

***Index Terms*—Data privacy, user profiling, personalized information systems, data generalization, data-perturbative mechanisms.**

I. Introduction

An estimated 4.1 billion people used the Internet in 2019 and, today, almost the entire world population stays connected to a mobile phone network [1]. The digital universe is also estimated to reach a capacity of 44 zettabytes (1 ZB $= 10^{21}$ bytes) this year. Phenomena such as *information overload*, *datafication* or *overloapped real and digital lives* accompany the fascinating and at the same time dizzying development of the Internet. New high impact technologies as the Big Data, IoT, 5G or Blockchain share scene with pressing global problems as the climate change, employment or recent pandemic.

At the same time, Schawb [2] highlighted in 2015 the future of the Internet as one of the ten most relevant issues facing the world. And from the social point of view, one of the most relevant challenges of this new industry are privacy concerns. In this sense, Schwab predicted that *debates about fundamental issues such as the impact on our inner lives of the loss of control over our data will only intensify in the years ahead*.

In this context, personalized information systems (PISs) are a clear example of the rise of the Internet and its risks to user privacy. Amazon, YouTube or Netflix are exponents of these systems, which have radically changed the way of accessing information with a high impact on our economy. PISs typically collect personal, behavioral data of their users over time to *profile* them and thus infer their interests or preferences and classify them. It is in this profiling process (defined by U.S. NIST [3] as *an operation or set of operations performed upon personally identifiable information (PII) that can include, but is not limited to, the collection, retention, logging, generation, transformation, use, disclosure, transfer, and disposal of PII*) where user privacy is compromised. However, because profiling is in fact what enables personalization, users of those systems are faced with a dilemma of great practical relevance: how to balance the trade-off between and privacy and personalization[1].

In the context of PISs, balancing this trade-off has been scarcely studied for local profile protection [4], that is, when the protection mechanism is put in place on the user side and the aim is countering the profiling carried out by those systems. This type of local protection is conventionally referred to as *hard privacy*, which, unlike *soft privacy* [5], [6], assumes users need not trust the service provider nor even the network operator, and hence, because they just trust themselves, it is their own responsibility to protect their privacy. Undoubtedly, hard privacy is aligned with the principle of *data minimization*, by which one would minimize the amount of data they

[1]We shall interchangeably use the terms *utility* and *personalization*.

would share with an information system without significantly degrading the expected functionality.

However, the literature of local mechanisms for profile protection can be classified essentially into two data-perturbation approaches: *forgery*, i.e., submitting false data; *suppression*, i.e., refraining from disclosing real, true data; and a combination of the two. In this work, we propose a novel category of data perturbation for profile-privacy enhancement, data *generalization*, which consists in replacing specific and sensitive data (e.g., a search query on AIDS treatment or a location at a hospital) with more general and thus less sensitive data, before sending them to the service provider in question.

We hasten to stress, however, that the idea of data generalization is by no means new. As a matter of fact, generalization is an old acquaintance of the field of statistical disclosure control (SDC) [7]. In SDC, generalization as well as forgery and suppression are applied to records of a database with the same aforementioned principle. Nonetheless, unlike SDC, our work investigates the application of generalization to protect a *different* data structure, namely, a user profile, which, like [8]–[23], we assume it is modeled mathematically as a probability mass function (PMF).

A. Contribution and plan of this paper

In this paper, we investigate data generalization as a hard-privacy mechanism that aims to protect PMF-based profiles against non-fully trusted PISs. Our mechanism assumes users are willing to generalize certain pieces of personal information while interacting with a PISs, in order to avoid being accurately profiled by this or, in general, by any privacy attacker capable of collecting such information. Following this simple principle, our approach may protect user privacy to a certain degree, without having to trust the system provider nor the network operator, but at the cost a loss in utility, a degradation of the quality of the personalized service.

Our first contribution is an architecture that implements this idea for a specific PISs, resource tagging, although it can be straightforwardly applied to any personalized service provider. The theoretical analysis of the trade-off between privacy and utility is our second contribution. We tackle this analysis in a systematic fashion, drawing upon the methodology of multiobjective optimization. We adopt a quantifiable measure of user privacy —the Shannon's entropy of the probability distribution of the user's data—, and formulate an optimization problem modeling the trade-off between privacy on the one hand, and on the other generalization rate as utility metric. Our theoretical analysis finds a closed-form expression of the critical generalization rate, i.e., the rate beyond which maximum privacy is attained. Our approach is then experimentally evaluated in a real-world application. Namely, we apply optimal data generalization to Foursquare and Brigthkite, two popular datasets from location-based social networks (LBSNs), and show, in a series of experiments, how our approach enables its users to enhance their privacy.

Section II explores some relevant approaches related to privacy and data-perturbative mechanisms in PISs. Section III describes our privacy-enhancing mechanism, some considerations about the user profile model and the adversary capabilities, and ultimately a formulation of the trade-off between privacy and generalization. Section IV presents in detail a possible practical implementation of our generalization mechanism, including some assumptions of our approach on a user profile. Section V presents a theoretical analysis of the optimization problem characterizing the privacy-generalization trade-off. In addition, this section shows a simples but insightful example that illustrates the formulation and theoretical analysis argued in the previous sections. Section VI presents an experimental evaluation of our technique in Foursquare and Brighkite datasets, including a discussion about it. Conclusions and future work are drawn in Section VII.

B. Practical implications and scope

Research in privacy protection against user profiling encompasses a wide range of methods and application scenarios. Notable examples include dynamic optimization in online mobile advertising to prevent tracking via temporary application usage [24], adversarial approaches to counter profile tracking through mouse cursor movements [25], and methods to safeguard data security through biometric sensors in user authentication [26]. Additionally, studies involve analyzing online behavioral models [27] and exploring advanced defense mechanisms in social networks and the Internet of Things (IoT). These mechanisms address contemporary challenges such as protecting user locations, especially for minors, and combating cyberbullying [28].

For the generalization of data in the context of PISs, we propose two applications that could be of great impact, given the large number of potential users that they would involve. Firstly, we consider the urban mobility scenario, where users demand services in exchange for sharing their location and/or their assessments of the functioning of public services, for example. We include in this scenario applications from a health point of view with an eye on the recent COVID-19 epidemic, as far as mobility and privacy are concerned. And secondly, we also contemplate the context of aviation, where travelers can be profiled, for example, based on their navigation path over any of the world's airports, or any other information of interest and usefulness to offer a personalized online service. We want to highlight that our experiments include generalization of spatial data, that is, on location coordinates (latitude and longitude), fundamental data on which the prolific field of location privacy and the popular location-based services are based.

Finally, we would like to note that no less notable is the versatility of the scope of privacy-enhancing technologies (PETs) based on data perturbation, as is the case of the mechanism here proposed. We believe that our contributions can significantly aid in protecting privacy against user profiling from two perspectives. Firstly, they are designed for privacy designers/engineers, aiming to equip them with a new tool for evaluating their systems' balance between user privacy and data utility for personalization. Secondly, these contributions empower users themselves. They enable users to protect their

privacy independently, without relying on third parties or sharing resources with others. Additionally, they assist users in determining the optimal amount of data perturbation needed to minimize the impact on the quality of personalized services they receive.

II. State of the Art

In the broad context of PISs, there are countless privacy-enhancing technologies (PETs) that allow the extraction and sharing of information while guaranteeing user privacy. According to [29], we can classify these technologies into five groups, namely, (a) basic anti-tracking technologies, (b) TTP-based approaches, (c) collaborative mechanisms, (d) cryptography-based methods from private information retrieval (PIR) and (e) data-perturbation techniques. The mechanisms that we investigate belong to this last class.

Focusing on this last group, data-perturbation techniques are a commonly used approach to block privacy attackers from attempting to accurately profile users, and based on obfuscating the information they explicitly or implicitly reveal when communicating. with a PIS. An illustrative example of this technique is sending false data, along with the user's genuine data. It is important to note that PETs are characterized by the level of trust that users place in the entities with which they communicate, and that, specifically, data-perturbation techniques usually adopt the untrusted model, assuming that any third party is a potential privacy attacker. Or what is the same, the data perturbation occurs on the user side, although this is not an obstacle for them to be combined with other more collaborative mechanisms.

If we take into account that personalization is the main objective of PISs and user profiling represents its main privacy risk, data perturbation poses the challenge of balancing the cost of functionality in the system and the utility of the data it implies, on the one hand, and protect the privacy of users, on the other.

In this section, we examine those works that aim to protect user profiles in PISs, and proceed depending on the type of perturbation technique applied: forgery, suppression, a combination of both. We would like to emphasize, though, that these two techniques (together with the one proposed in this work based on generalization) can also be found in the field of SDC. However, the object of protection of the works analyzed in this section is a user profile, typically modeled as a PMF, whereas in SDC, those techniques aim to protect a whole database of records.

A. Forgery

The idea behind query forgery is simply based on adding fake queries (or keywords) to the original ones. This approach allows users to obfuscate their profile, protecting them from precise profiling by privacy attackers, and avoiding the need to trust potentially harmful third parties.

There are different alternatives based on the falsification of queries such as the system for private Web browsing called PRAW [30]–[33]. This system protects the privacy of a group of users who access the Web through shared access. Assuming that users have logged in to a website and are therefore identified, while the attackers' efforts focus on profiling the group, PRAW hides the users' real profiles by generating false navigation traces in order to preserve their privacy.

Inspired by the same methodology of false query injection, in [34], the authors propose a generator of real and false queries based on complementary probabilities that are only available on the user side, and therefore assumed to be unknown to attackers of privacy.

And a popular web browsing plugin that implements query forgery using different strategies is TrackMeNot [35]. An integrated keyword dictionary, which is fed with disparate sources of information, is the source used to generate the false queries that are forwarded to the server. Communication can be simulated in bursts, as if it were human web searches, or in time intervals. However, the vulnerability of TrackMeNot to certain attacks based on semantics or the time between false queries to distinguish them from the real one is argued in [36].

Another proposal for profile obfuscation in software form is GooPir [37]. Its operation is based on mixing genuine query keywords with false keys that prove similar use (frequency). The result is sent in batches of keywords to the web search engine, so that attack attempts on the user's query profiles are hindered at the precise moment. However, [38] presents a criticism of GooPir in the sense that its strategy does not prevent an attacker from being able to calculate correlations between keywords from different batches and finally infer the user's real interest.

Lastly, in any case it is essential to take into account the implicit handicap of adding false queries, which is clearly the traffic overload. This circumstance implies considering a trade-off between privacy and added traffic. Precisely, in [12] the authors investigate theoretically this balance in the field of information retrieval with a mathematical model with the aim of optimizing the percentage of falsified queries and the privacy of users.

B. Suppression

Supression is the opposite alternative to introducing (false) activity into a user profile and is a perfectly viable perturbative technique. This is demonstrated by the authors in [13] on the scene of the semantic Web. The removal of labels, a process with associated costs in terms of resources, allows the user to improve their privacy although in a limited way, in compromise with the semantic degradation of the data. The Shannon entropy of the perturbed profile and the percentage of labels that the user is willing to delete are, respectively, the privacy and utility metrics on which the authors study their optimal balance through convex optimization in . Along this line, [16] presents an interesting application of label suppression in the context of resource recommendation and parental control. In this case, the impact of the perturbation is evaluated both in terms of costs associated with data degradation and in the precision of the assignment of one or another predefined parental control policies.

C. Combined Forgery and Suppression

Combining forgery and suppression is a strategy used in the scenario of personalized recommender systems, such as Movielens[2]. In practical terms, coupling both techniques allows users to send false ratings and/or not rate elements that are of interest to them. In this subfield of PISs, the trade-off between privacy protection and utility of user profiles has been the subject of research in [14]. More specifically, in [17], the authors contribute a closed solution to the problem of optimal and simultaneous forgery and suppression of ratings, which they evaluate on the real-world Movielens dataset.

III. Privacy Protection via Data Generalization

Based on the hard privacy assumptions [6], the optimal generalization is a privacy mechanism that is intended to prevent privacy attackers from profiling users on the basis of the data they share with the system. Conceptually, our approach protects user privacy to a certain extent, generalizing those data that make a user profile show a bias towards certain categories of interest. From a practical perspective, our profiling generalization technique is conceived to be deployed as a software application that runs on users' local machines. The software implementation is then responsible, on the one hand, for warning the user when their privacy is being compromised and, on the other, for helping them decide which data should and should not be generalized. Consequently, our approach ensures user privacy to some extent without having to rely on an external entity, but at the cost of some local processing expenses and, more importantly, the information loss incurred by the generalization of data.

In this section, we first propose a user profile model and next describe our assumptions about the adversary capabilities. Then, we justify a quantitative measure of the privacy of this profile. These considerations lead us to formulate the problem of choosing a optimal generalization strategy as a multi-objective optimization problem that takes into account both privacy and generalization rate.

A. User Profile Model

In our scenario of PISs, we assume there is a computerized system that builds profiles from the activity data the system collects over a time period. However, in the construction of those profiles, users directly regulate their privacy and, therefore, the data they finally share with the system. For this reason, the user must make a decision: share some data without manipulating/altering them, or transmitting a perturbed and more general version of those data. Clearly, depending on the level of generalization applied for each collected data, the profile recorded in the server will resemble, to a greater or lesser extent, the genuine, accurate profile of an individual. In this work, we shall refer to these two profiles as the *actual individual profile* and the *apparent individual profile*, and denote them by q and t, respectively.

[2]https://movielens.org/

Accordingly, define q as the probability mass function (PMF) of the authentic data of a particular individual, and p as a target distribution that is considered *privacy insensitive*. That is, if an individual's distribution, built from his/her recorded data, was observed to be p by the service provider, then the individual would accept there is no privacy risk in the profiling of his/her activity.

B. Adversary Model

Our technique is built on the principle of generalization data. Under this principle, a user may wish to generalize some pieces of information to enable the resulting user profile t, as observed from the outside, to approach the *uniform profile*, which we denote by u.

Bearing in mind this consideration and the user profile model described in Section III-A, we assume an adversary model in which users submitting individual information (either generalized or not) are observed by a passive attacker whose main aim is to profile them based on the observed information.

Last but not least, we also assume that the privacy attacker is unable to discern whether a particular user is adhered to the proposed privacy strategy, and therefore cannot estimate their generalization rate.

C. Generalization Model

We shall adopt the same notation for vectors used in [39]. Specifically, we delimit vectors and matrices with square brackets, with the components separated by space, and use parentheses to construct column vectors from comma separated lists.

We model individual private data (e.g., location tags, music tastes, GPS coordinates, ratings and bookmarks) as a sequence of random variables (r.v.'s) taking on values in a common finite alphabet of categories, in particular the set $\mathscr{X} = \{1, \ldots, n\}$ for some integer $n \geqslant 2$. In our mathematical model, we assume these r.v.'s are independent and identically distributed. This assumption permits us to represent the profile of an individual by means of the probability mass function (PMF) according to which such r.v.'s are distributed, a model that is widely accepted in the privacy and security literature [9], [12], [40]–[42]. Conceptually, we may interpret a profile as a histogram of relative frequencies of individual data within a predefined set of categories of interest.

Intrinsic to data generalization is the existence of a hierarchy of concepts or taxonomy. In this work, we shall denote by d the number of hierarchy levels other than the bottom-level, and assume that the highest-level category (root) can be reached from any bottom-level category (leaf) in exactly d jumps.

For $r = 1, \ldots, d$, we denote by $g^r = (g_1^r, \ldots, g_n^r)$ a *generalization strategy* of level r, which is a tuple specifying the percentage of user data that is generalized to that level and recorded as such. Accordingly, we define $G \in \mathbb{R}_+^{n \times d}$ as the matrix whose r-th column is g^r.

We model the way profiles are updated via the set of matrices $U^r \in \mathbb{R}_+^{n \times n}$ for $r = 1, \ldots, d$. Intuitively, when the proposed mechanism (see Sec. IV for further details)

generalizes a certain piece of data (e.g., a location) into a higher-level category h, the mechanism

i. implicitly refrains from recording that data, and

ii. updates its knowledge about the individual's profile on *all bottom-level categories* falling below h.

The update can be made uniformly across all such bottom-level categories, or proportionally, according to some distribution. For simplicity, we shall assume the former case.

For a 1-level hierarchy, we model the profile that results from optimal generalization (i.e., the apparent profile) as

$$t = q - g^1 + U^1\, g^1. \tag{1}$$

We immediately note that generalization can be regarded as a combination of suppression and partial forgery. In the case of the hierarchy depicted in Fig. 1, where $n = 10$ and $d = 1$, the matrix U^1 takes this form:

$$U^1 = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

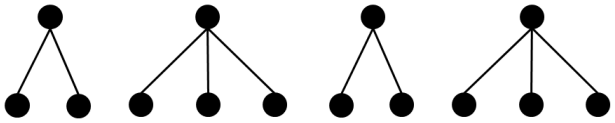


Fig. 1: Example of 1-level hierarchy.

Let

$$V^r = U^r - I_n,$$

where I_n is the identity matrix of size $n \times n$. It is interesting to note that V^r matrices are finally diagonal block matrices formed by the direct sum of negative centering matrices. Accordingly, in the general case when $d \geqslant 1$, the apparent profile is modeled as

$$t = q + \sum_{r=1}^{d} V^r\, g^r. \tag{2}$$

Note from expressions (1) and (2) that we are dealing with a combination of suppression and proportional forgery on the actual profile, equivalent to a translation and centering in terms of matrix algebra.

In this work, we shall assume all the data an individual generates can be classified into a bottom-level category of that taxonomy. In other words, we consider that the data collected by both our proposed privacy mechanism and the personalized service provider always refers to specific information about an individual; the aim of the proposed mechanism is precisely generalizing those individual data.

The proposed generalization mechanism captures the utility loss incurred by generalizing individual data through a cost matrix of dimension $n \times d$

$$C = \begin{bmatrix} c_{11} & \dots & c_{1d} \\ c_{21} & \dots & c_{2d} \\ \vdots & & \vdots \\ c_{n1} & \dots & c_{nd} \end{bmatrix},$$

where the entry $c_{ir} \geqslant 0$ reflects the impact on utility due to generalizing a private data (e.g., a location) that belongs to the bottom-level subcategory i, into a category of the r-th level of the hierarchy.

For a single individual, we quantify the effectiveness of the protection mechanism in generalizing user data, through the accumulated total cost

$$\sum_{r=1}^{d} c_{\cdot,r}^{\mathrm{T}}\, g^r,$$

which each individual or user would like to upper bound. For simplicity, we shall assume the same bound is applied to all individuals within the population, and denote it by γ.

D. Trade-Off between Privacy and Generalization Rate

We quantify an individual's *privacy risk* generically as

$$\mathcal{R} = f(t, p),$$

where $f(t, p)$ is a *privacy function* that measures the extent to which the individual is discontent when the apparent profile is t and the target profile they would like to look like is p. As a first approximation, we may consider the target profile to be the uniform distribution u, and the privacy function to be the Shannon's entropy of the apparent profile,

$$\mathcal{P} = \mathrm{H}(t),$$

and refer to it consistently as *privacy gain*, rather than privacy risk. We use Shannon's entropy to reflect the intuition that an attacker will be able to compromise user privacy as long as the apparent user profile diverges from the uniform profile. Recall [43] that the entropy of a PMF t is defined as

$$\mathrm{H}(t) = -\sum_{i=1} t_i log_b(t_i),$$

where b is the base of the logarithm used. Common values of b are 2, e and 10. In those cases, the units of entropy are *bit*, *nat* and *dit*, respectively. For simplicity, we shall use natural logarithms throughout the paper and refer to $\log_e$ as $\ln$, particularly because all bases produce equivalent optimization objectives.

Consistently with this privacy entropic measure, now we define the privacy–generalization function, or equivalently, the

optimal privacy-utility trade-off between privacy and generalization for an *individual*

$$\mathcal{P}(\gamma) = \max_{\substack{G \succcurlyeq 0 \\ G\,\mathbf{1} \preccurlyeq q, \\ \sum_{r=1}^{d} c_{\cdot,r}^{\mathrm{T}}\, g^{r} = \gamma}} \mathrm{H}(t), \tag{3}$$

which formally expresses the intuitive reasoning behind data generalization: the higher the data generalization rate γ, the higher the uncertainty in terms of the entropy of the apparent distribution, and the higher user privacy.

IV. User-Side Architecture

We dedicate this section to describing, adapting a basic high-level architecture proposed in [44], an eventual and practical implementation of our optimal data generalization mechanism. We know that the main purpose of our solution is for users to consciously protect their privacy against identification attempts by any attacker. And to do this, we assist them in deciding what proportion of data they should generalize at the cost of losing the minimum functionality in the service they receive. From this point of view, we want to highlight that we also conceive our proposal as a *decision support system*.

Our approach is designed to be integrated into an application that the user already has installed on their computer, for example, a web browser plug-in. Inspired by the principles of hard privacy, the architecture we present is based on the untrusted model, eliminating the need for users to trust third parties to protect their data, beyond the software installed on their own machine. It is for this reason that we call it *user-side architecture*.

The operation that the application must carry out is straightforward. When the user's privacy is at risk, the system launches an alarm and then recommends them what data should be generalized to deal with the threat. Hence, it is the user who has the last word.

However, before detailing the main functional components of our design, we must specify how a user's profile could be obtained locally in an application that implements our technique. To do this, we base ourselves on three assumptions about said profile.

1) Common knowledge. First, We assume that both instances, the software application and the eventual privacy attackers, operate on an identical predefined taxonomy of categories of interest and therefore obtain, based on their categorization algorithms, the same user profile. This assumption is considered valid to the extent that we are dealing with sets of standard and generalized categories.
2) Profile inicialization. Second, we assume that to decide whether to generalize a particular category or not, our approach needs an initial user profile. This circumstance can be taken into account through a training period prior to the implementation of the architecture we propose. Exclusively during this previous phase, the user will explain her interests since under this assumption, an attacker could know her real profile.
3) Long-term profile. And additionally, we assume that the user's profile is not subject to frequent changes, in line with the so-called long-term profiles [45]. The profile acquires stability after the initialization phase previously noted, when the user has shared a significant number of elements. Still, we must recognize that, in practice, user interests can vary significantly over time and, therefore, our solution must take this into account.

Figure 2 describes, through a block diagram, the software architecture that we propose as an application. It is made up of a series of modules that interact locally and/or with the system, so that each of them performs a specific function based on the parameters it receives. Although our solution can work with *any* type of generalizable data, the figure shows the case of resource tagging on the Web, that is, when users' private data are tags.

From a general perspective, the figure shows a user interacting with a single, simple PIS, and more specifically, with an *hierarchical tagging system*. It is an entity that, in exchange for personalized information of interest, stores and provides users with elements of information (for example, music, videos and web pages, points of interest or geographical coordinates) and their corresponding associated tags, which can form part of a hierarchical taxonomy. Below we provide a functional description of the modules that make up this architecture.

Web browser. Unlike the rest of the modules, we assume that the web browser can be previously installed on the user's computer. This is an external element to our generalization application, which we understand as a complement to the first. The browser is responsible for the user's communication with the PIS, entities that feed each other information. Thus, a user downloads through his or her browser that content (e.g. images, web pages, etc.) that will be tagged according to his or her interests, including the tags that other users have published in the tagging system. And in the same way, the browser sends the labels proposed by the user to the PIS. Meanwhile, all data retrieved by the browser (e.g. metadata) is passed to the *context analyzer* module to process the information.

Context analyzer. The purpose of this module is to assist the *category extractor* module in deciding which user profile category should be updated (generalized). This process could be carried out using the vector space model [46], as is normally done in the field of information retrieval, to represent Web pages as tuples containing their most representative terms. For example, term inverse document frequency (TF-IDF) could be applied to calculate the weights of each term that appears on the web page that includes the item to be categorized. Subsequently, taking some of the most weighted terms from the tuple, it would send them to the *category extractor* module.

Category extractor. This element plays a crucial role in classifying the tags that the user submits to the PIS within a predefined set of categories. In certain cases, these categories are provided by the labeling system, as seen on platforms such as Amazon, but they can also be acquired through specialized

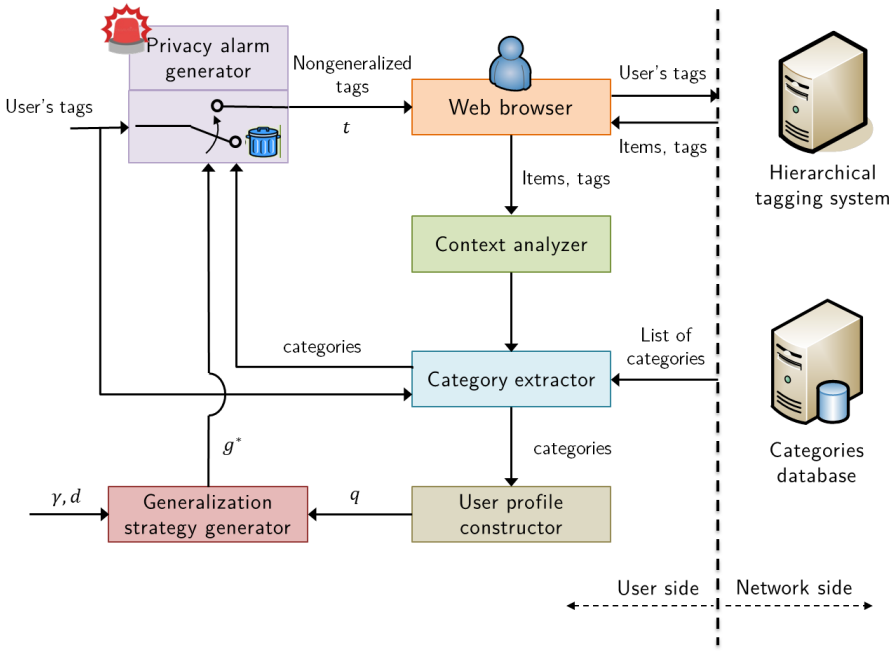


Fig. 2: Block diagram of the proposed architecture given a hierarchical taxonomy with labels arranged into a tree structure.

databases, such as the Curlie Directory [3]. During this process, the label suggested by the user and the contextual information provided by the *context analyzer* module are integrated. The result, presented in the form of categories, is transmitted to the *user profile constructor* and *privacy alarm generator* modules.

User profile constructor. This module generates the user profile. In specific terms, it receives the categories associated with the tags sent by the user and, accordingly, estimates and/or updates their profile. As mentioned above, our proposed architecture assumes that, when estimating the histogram, the relative frequencies of activity are sufficiently stable once the user has generated a significant number of labels. A crucial aspect that a practical implementation of this module must take into account is profile initialization. One option could be to start this profile at zero [47]. On the other hand, an approach based on the principle of maximum entropy would choose to use the uniform distribution. Importantly, this module remains active even when the user explicitly declares her profile. Since the profile indicated by the user may not accurately reflect their online behavior, our architecture may decide, after the training phase, to replace it with the profile implicitly inferred from their interaction activity (tagging) with the system.

Generalization strategy generator. This module is responsible for the user's privacy and therefore we consider it the central element of the architecture we propose. Equipped with the implementation of our optimal generalization mechanism, and based on two input parameters, the real user profile q, the number of hierarchy levels other than the bottom-level d and the generalization rate γ, which represents the percentage of data to be generalized, the result of its calculations is the optimal tuple of generalized data g^*. For example, the component g_i^* refers to the percentage of data that is suggested to be categorized into the top-level category i. In Section III-D, we provide more detailed specification of this module.

Privacy alarm generator. The task of this module is to alert the user about possible violations of their privacy. Whenever the user shares their personal information through tags, this module waits for the *category extractor* module to send the category corresponding to said tag, represented by the index i. And when it receives the tuple, it executes the following actions. A privacy warning with probability g_i is generated, notifying the user. It is the user's responsibility to decide whether or not to generalize the data in case of alarm activation. And otherwise, our application will not identify any privacy threats and will transmit the information to the *web browser*.

V. Theoretical Analysis

In this section, we shall analyze one of the main and fundamental properties of the privacy-generalization function (3) defined in the previous section. Our theoretical analysis only considers the case when all given probabilities are strictly positive:

$$q_i > 0 \quad \text{for all} \quad i = 1, \ldots, n. \tag{4}$$

This assumption will be properly justified in Section V-A. We shall suppose further, now without loss of generality, that

$$q_i \leq \cdots \leq q_n \quad \text{for all} \quad i = 1, \ldots, n. \tag{5}$$

Before proceeding with the mathematical analysis, it is immediate from the definition of the privacy-generalization

[3]https://curlie.org/

TABLE I: Description of the variables used in our notation.

Symbol	Description
n	Number of interest categories
d	Number of hierarchy levels
m	Number of top-level categories
n_k	Number of bottom-level categories by each top-level category
g^r	A *generalization strategy* is an n-tuple with the percentage of observed data that is generalized at level r
q	The *actual* user profile is the genuine profile of interests
t	The *apparent* user profile is the perturbed profile, as observed from the outside, resulting from the generalization of certain percentage of observed data
u	Uniform profile across the n categories
V^r	The *generalization matrix* at level r
C	The *cost matrix* on wich the proposed generalization mechanism captures the utility loss incurred by generalizing individual data
$\mathrm{H}(t)$	User privacy is measured as the *Shannon's entropy* of the apparent user profile
γ	The *generalization rate* is the percentage of observed data that the user is willing to generalize
$\mathcal{P}(\gamma)$	Function modeling the privacy–generalization trade-off
γ_{crit}	The *critical generalization* is the generalization rate beyond which the privacy–generalization function attains its maximum value or critical privacy

function that its initial value is $\mathcal{P}(0) = \mathrm{H}(q)$. The notation used throughout this section is summarized in Table I.

A. Critical Generalization

The following theoretical property confirms the intuition that there must exist a generalization rate beyond which critical privacy is achievable, in the sense that the privacy–generalization function attains its maximum theoretical value by groups of top-level categories, that is, $\mathcal{P}(\gamma) = \phi \leq \ln(n)$ [4]. This *critical generalization* is

$$\gamma_{crit} = 1 - \sum_{k=1}^{m^d} n_k q_{1k}, \tag{6}$$

where m^d is the number of category sets at top-level of the hierarchy and according to the labeling assumption (5). Our purpose is to formulate and prove a theorem that captures this property. It is important to note that this property also provides an upper bound on the value of $\mathcal{P}(\gamma)$.

*Theorem 1 (**Critical generalization**):* For all $\gamma \in [0, 1)$, if $\gamma \geq \gamma_{crit}$, then $\mathcal{P}(\gamma) = \sum_{k=1}^{m} n_k \, \mathrm{H}(\bar{q}_k)$. Conversely, if $\gamma < \gamma_{crit}$, then $\mathcal{P}(\gamma) < \sum_{k=1}^{m} n_k \, \mathrm{H}(\bar{q}_k)$.

Proof: First, suppose that at the maximum value of $\mathcal{P}(\gamma)$ the apparent profile t is constant (uniform) for groups of top-level categories. In the case of a single hierarchy level ($d = 1$), by algebraic manipulation of Eq. (1) we can express the resources g in the form $g_{ik} = g_{1k} + q_{ik} - q_{1k} \; \forall i > 1, k$. Adding all the terms we arrive at an expression for the generalization rate, that is, $\gamma = 1 + \sum_{k=1}^{m} n_k g_{1k} - \sum_{k=1}^{m} n_k q_{1k}$. For γ to reach its highest value within the interval $[0, 1)$, the first component of the resources of each group of high-level categories will necessarily have to be zero, that is, $g_{1k} = 0$, $k = 1, \ldots, m$, thus obtaining the desired Eq. (6).

[4]Note that with only one set of top-level categories ($m = 1$), it follows that $t = u = 1/n$ and $\mathcal{P}(\gamma) = ln(n)$.

Furthermore, in all cases we can express the optimal resource strategy for the critical generalization ratio in the form $g_{ik} = q_{ik} - q_{1k}$ for all $i > 1$ and k. In this way, the apparent profile t will have a uniform value for each group of top-level categories, that is, $t_{ik} = \bar{q}_k$, $k = 1, \ldots, m$, understanding $\bar{q}_k$ as the average of actual user profile at the k-th group of categories. Then, the critical value of the privacy–generalization function is $\mathcal{P}(\gamma) = \sum_{k=1}^{m} n_k \, \mathrm{H}(\bar{q}_k)$. Following the same reasoning, the particular case ($d = 1$) is extensible to the general case. ■

*Corollary 2 (**Privacy bound**):* The critical privacy $\mathcal{P}(\gamma) = \sum_{k=1}^{m} n_k \, \mathrm{H}(\bar{q}_k)$ is upper bounded by $\ln(n)$.

Proof: From Gibbs' inequality we know that H_n attains its maximum value when all probabilities are equal, i.e., $\mathrm{H}_n(q_1, \ldots, q_n) \leq \mathrm{H}_n(1/n, \ldots, 1/n)$. In that case, $\mathrm{H}_n(1/n, \ldots, 1/n) = \ln(n)$. In turn, according to the structure of our problem we have $-\sum_{k=1}^{m} \sum_{i=1}^{n_k} 1/n \ln(1/n) = \sum_{k=1}^{m} n_k \, \mathrm{H}(1/n) = ln(n)$. It follows then that $\mathcal{P}(\gamma) = \sum_{k=1}^{m} n_k \, \mathrm{H}(\bar{q}_k) \leq \sum_{k=1}^{m} n_k \, \mathrm{H}(1/n) = \ln(n)$. ■

Comparison with forgery and suppression. The analytical characterization of the critical generalization rate is a fundamental result: we can analytically determine the amount of perturbation needed to achieve maximum protection, and based on this, figure out whether or not a user can achieve it more or less easily.

Although obviously the critical rate depends on each profile, a natural question one would ask is: how is this critical rate in relation to state-of-the-art perturbation techniques, namely, forgery and suppression. We aim to shed some light into this question next.

We conduct our first experiment for some concrete user profile, evaluating the state-of-the-art techniques optimal forgery [12] and optimal suppression [15] [19], versus the proposed optimal generalization mechanism. Together with [12], henceforth denoted `forgery v.1`, we also assess a variation thereof (`forgery v.2`), which models the apparent profile as follows: $t = (q + r)/(1 + \rho)$, where ρ is the perturbation rate and r a forgery strategy analogous to our g. This model is in contrast with the original formulation of optimal forgery ([12]), where $t = (1 - \rho)\, q + \rho\, r$.

Fig. 3 shows the results for the user profile

$$q = (0.02, 0.03, 0.04, 0.05, 0.07, 0.10, 0.12, 0.15, 0.17, 0.25)$$

and the hierarchy "3-level toy28" of Fig. 5(e). In the figure, we can observe that, *for this specific profile*,

$$\rho_{\text{crit}} < \gamma_{\text{crit}} < \sigma_{\text{crit}},$$

which means that forgery can provide the highest level of protection with a smaller perturbation rate than generalization and suppression would do. In the figure, we would like to note that, unlike generalization, suppression and forgery rely on bottom-level categories to perturb profiles (which is denoted with the label "0-level" in the figure). The fact that there is no hierarchy of categories in suppression and forgery implies, on account of Theorem 1, that the privacy level attained by

these two techniques will never be smaller than that offered by generalization.

Because the results reported in Fig. 3 are highly dependent on the chosen profile, our second experiment contemplates 608 random profiles of dimension $n = 16$ over some possible high-level hierarchies. More specifically, for each of these random profiles, we show in Fig. 4 the critical rates of generalization, suppression and forgery in the ordinate, and the difference $q_{16} - q_1$ in the abscissa. Our choice for the abscissa is justified by the fact that $\rho_{\text{crit}} = 1 - 1/n\, q_n$ [12] and $\sigma_{\text{crit}} = 1 - n\, q_1$ [15], which, together with Eq. (6), imply q_1 and q_n are the only profile components affecting the three critical rates.

Hierarchy-wise, from Eq. (6) it follows that γ_{crit} depends just on the highest level. To explore the effect of the number of high-level categories in generalization, data points (i.e., profiles) in Fig. 4 with darker color reflect profiles with higher number of such categories.

Fig. 4 shows the superiority of generalization in terms of critical rates with respect to the state-of-the-art perturbation techniques. For the random profiles generated, optimal generalization largely achieves the maximum privacy level $\sum_{k=1}^{m} n_k\, \mathrm{H}(\bar{q}_k)$ at significantly lower perturbation rates than forgery and suppression. In Sec. VI we shall examine deeper this question with real profiles.

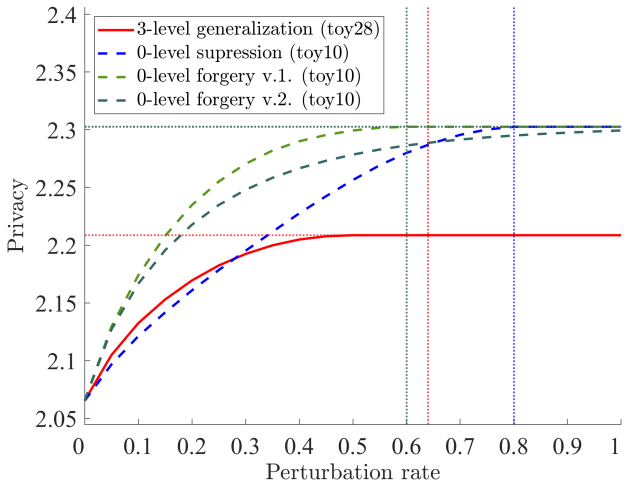


Fig. 3: Privacy, measured as the Shannon's entropy of a user's apparent profile, vs perturbation rate, for 3-level generalization, suppression and two versions of the forgery technique.

B. Numerical Example

In this section, we show some numerical results for a simple but insightful example that will illustrate the formulation presented in Section III-C and III-D and the theoretical analysis argued in Section V. Throughout this subsection, all results correspond to a same artificial user. The experimental analysis of our privacy-enhancing mechanism in a real-world application is presented later in Section VI.

In this practical example, we shall consider $n = 10$ bottom-level categories and assume that the user distribution is again $q = (0.02, 0.03, 0.04, 0.05, 0.07, 0.10, 0.12, 0.15, 0.17, 0.25)$,

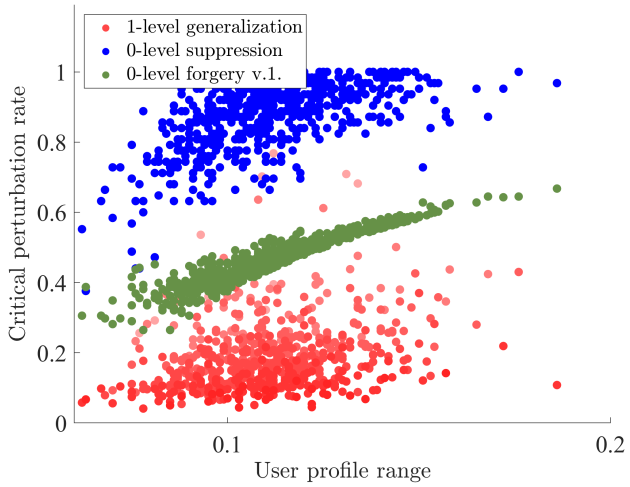
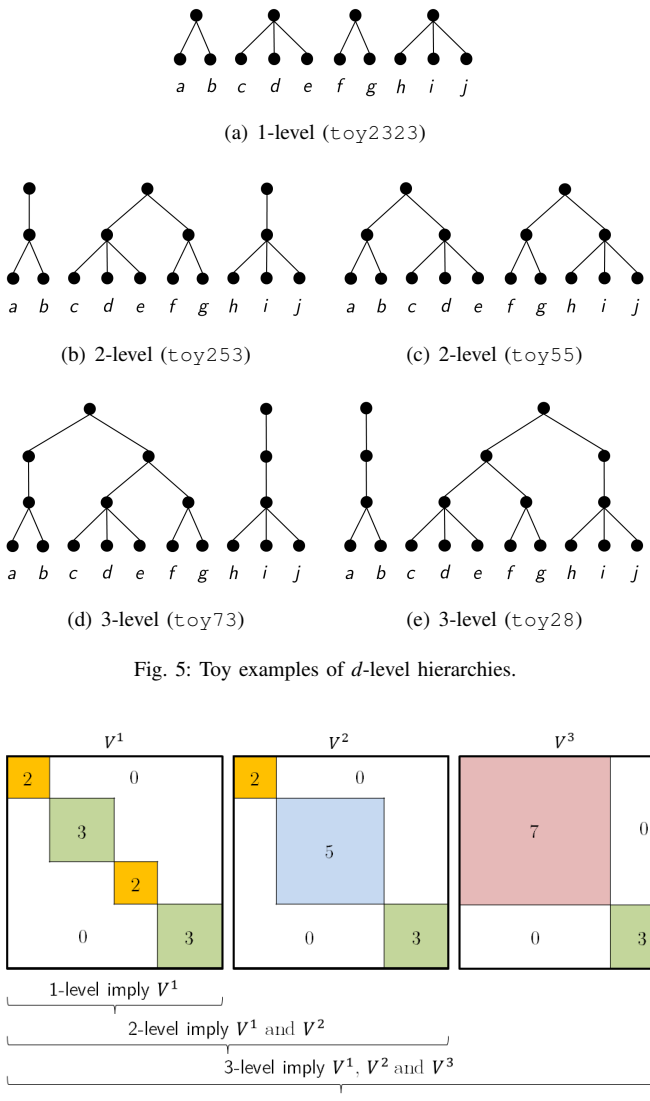


Fig. 4: Critical perturbation rates of 608 randomly-generated, 16-dimensional profiles vs the difference $q_{16} - q_1$, for optimal generalization, suppression and forgery v.1. Each data point reflects a profile, and darker profiles in generalization imply larger number of high-level categories.

thus fulfilling both the positivity and the labeling assumptions 4 and (5). For simplicity, we also assume that the cost matrix C is the all-ones matrix. Furthermore, on the same user profile, we consider five possible generalization hierarchies with levels $d = 1, 2$ and 3, which we illustrate in Fig. 5. Note that the examples evolve from the lower prelevel example, starting with the same basic 1-level example, namely `toy2323`. In Fig. 6 we can observe the structure of the generalization matrices V^1, V^2 and V^3 in order to form in each case the corresponding Eq. (2) from the 1-level `toy2323`, 2-level `toy253` and 3-level `toy73` examples. Notice how, at each level, the generalization matrix has as many diagonal blocks as groups of categories that have been defined for that level. Furthermore, the dimension of each square block corresponds to the number of lowest level categories that the group of categories contains.

To solve numerically the optimization problem (3) we used CVX, a package for specifying and solving convex programs [48], [49] and MOSEK [50]. Both have been run on Matlab software (Matlab R2021 9.10.0.1602886 64-bit win64) with an Intel CoreTM i3-2370 2.4 GHz CPU, 4Gb RAM in a Windows 10 64-bit operating system.

From the point of view of the user profiles, in Fig. 7 we represent the apparent profile of the user from the 3-level `toy73` example, for different values of the generalition rate γ. When $\gamma = 0$, no perturbation takes place and the apparent profile t represented in Fig. 7(a) actually corresponds to the genuine user profile q. According to the reasoning behind the optimal generalization strategies previously described, the higher γ, the more uniform is the resulting apparent profile. As illustrated in Fig. 7(d), the maximum level of privacy is attained precisely for $\gamma = \gamma_{crit} = 0.41$, when the apparent profile is completely uniform by the two top-level categories and therefore, following the expressions obtained in the Theorema 1,



(a) 1-level (toy2323)

(b) 2-level (toy253)

(c) 2-level (toy55)

(d) 3-level (toy73)

(e) 3-level (toy28)

Fig. 5: Toy examples of d-level hierarchies.

Fig. 6: Hierarchical matrices from toy2323, toy253 and toy73 examples.

$t^* \simeq (0.06, 0.06, 0.06, 0.06, 0.06, 0.06, 0.06, 0.19, 0.19, 0.19)$ and $\mathrm{H}(t^*) \simeq 2.1463$.

All this information is also captured in Fig. 8, where we plot the privacy-generalization function (3), that is, the function modeling the optimal trade-off between privacy and utility, the latter being measured as the percentage of data generalized by the user. However, in the figure we collect the information for the five hierarchy examples that we use on the same user.

From this point of view, the first thing we observe is that in all cases, the function $\mathcal{P}(\gamma)$ behaves non-decreasing and quasiconcave, as we pointed out at the end of our theoretical analysis. Starting from a common initial value, namely $\mathcal{P}(0) = \mathrm{H}(q) = 2.0652$, and depending on the rate of generalization γ, the optimal trade-off grows with different intensities until it reaches critical value of the rate, namely γ_{crit}, and following the calculations of 6, at which privacy growth stalls. In this sense, the best increase of 7% is obtained when the user categorizes their interests under the 3-level toy28, so that $\mathrm{H}(t^*) =\simeq 2.2086$ although at the cost of a critical rate $\gamma_{\text{crit}} = 0.64$.

A final remark is the influence of the hierarchy level d on user privacy. We observe that the higher the hierarchy level, the greater the privacy.

VI. Experimental Analysis

In this section, we will discuss the extent to which our technique enables users to enhance their privacy in a personalized real-world system. Our analysis also looks at the impact that data generalization has on information loss using a measure of utility that we call the generalization rate, namely γ.

A. Datasets

We applied the proposed technique to Foursquare [51] and Brightkite [52], two real-world datasets collected from location-based social networks (LBSNs) which are well-know by the scientific community for data mining tasks in the field.

The Foursquare dataset employed in our experiments was originally used for studying the spatial-temporal regularity of user activity in LBSNs. This publicly available dataset [5] contains 227.428 check-ins collected in New York City for about ten month, from April 2012 to February 2013. Each check-in is associated with its time stamp, its GPS coordinates and its semantic meaning represented by fine-grained venue-categories [6].

The Brightkite was once a LBSNs provider where users shared their locations by checking-in. From the friendship network, this publicly available dataset [7] was collected in New York City [8] and contains a total of 123.558 check-ins of these users over the period of April 2008 to October 2010.

In Fig. 9 we present a geographic map of each dataset.

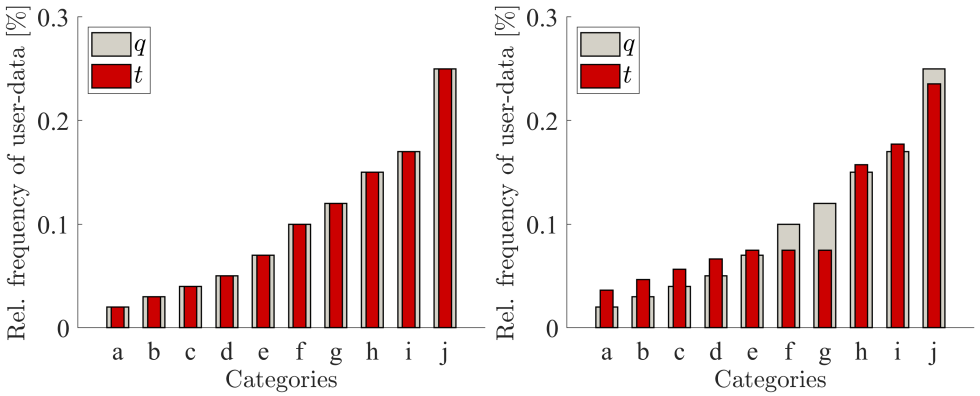
B. Hierarchical Categorization

Our model represents user profiles as normalized histograms of relative frequencies of individual data within a set of its categories of interest. To evaluate our technique, we need to adapt the datasets and first categorize the data. In this way, we define categorization-based experiments depending on the selected attributes that we will take by categories.

For what we call *semantical* experiments, we shall rely on the labeled categories associated to the venues provided by the LBSN in the case of Foursquare. However, this dataset only provides a hierarchy of two-level depth ($d = 2$), as illustrated by the example in Fig. 10. For example, the interest of a user who checks in when visiting a "cupcake shop" can be presented in a more general way as "dessert shop" or even more as "food".

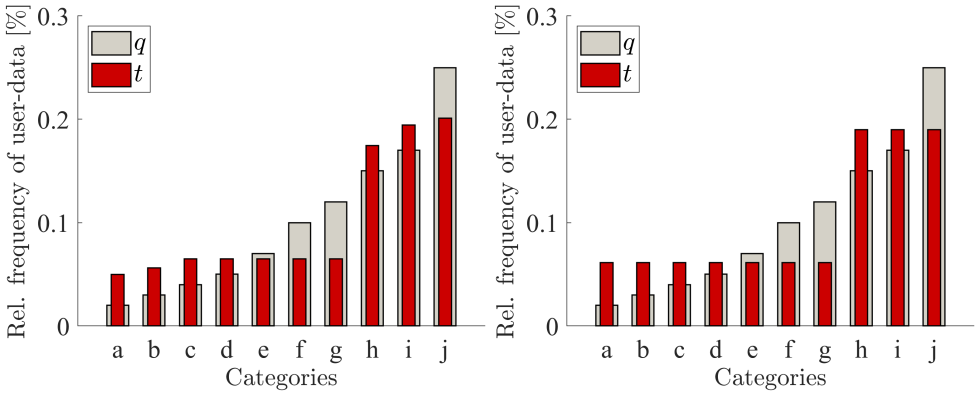
On the other hand, for the so-called *spatial* experiments, we shall use the Cartesian coordinates of the user locations in case of both datasets. These coordinates are expressed

[5] https://sites.google.com/site/yangdingqi/home/foursquare-dataset
[6] https://developer.foursquare.com/docs/categories
[7] https://snap.stanford.edu/data/loc-brightkite.html
[8] NYC area between the latitude coordinates (40.5509 and 40.9883) and longitude (-74.2748 and -73.6838).



(a) $\gamma = 0$, $\mathrm{H}(t) = 2.0652$

(b) $\gamma = 0.1367$, $\mathrm{H}(t) = 2.1253$



(c) $\gamma = 0.2733$, $\mathrm{H}(t) = 2.1434$

(d) $\gamma = \gamma_{\text{crit}} = 0.4100$, $\mathrm{H}(t) = 2.1463$

Fig. 7: Actual and apparent profiles for different values of γ from the 3-level `toy73` example.

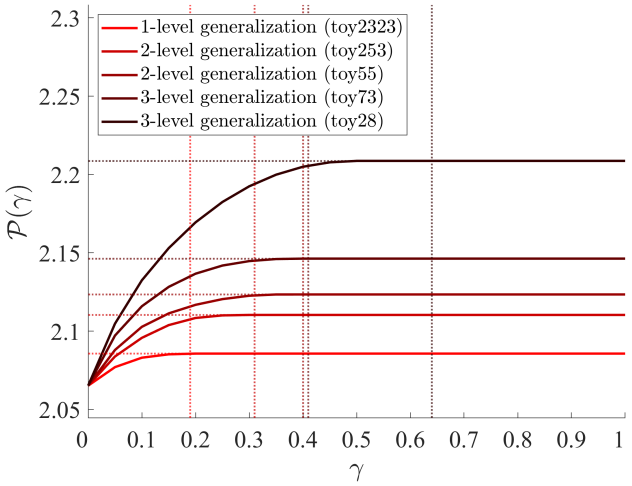


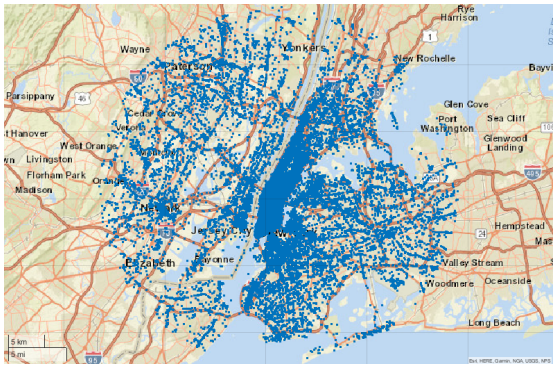
Fig. 8: Privacy function vs generalization rate γ. Vertical dashed line indicate critical generalization rate γ_{crit} from each toy.

in kilometers and have been translated from latitude and longitude coordinates using the center of the New York City region as a reference and the Haversine formula [53]. To obtain the corresponding hierarchy, we divided the area in four quadrants recursively until the level of granularity (or depth) that we want, and categorize each pair of user coordinates means of the nearest center search to each quadrant. In this sense, we must take into account that the spatial hierarchy will always form groups of categories made up of a maximum of four lower level categories. In Fig. 11 we illustrate the recursive partition idea. For example, in the case of having defined a hierarchy two levels deep, the coordinates of a user who shares their location in the upper westernmost part of the New York City area can be categorized from least to most general with the labels "111", "11" and "1".

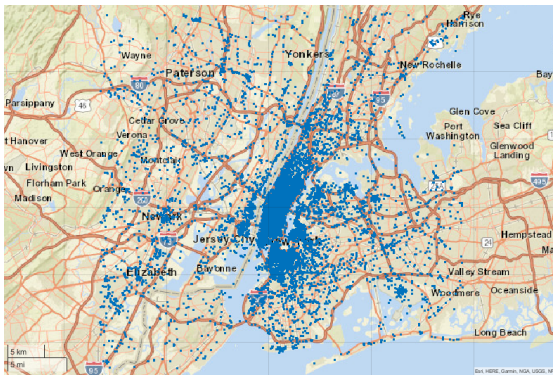
Then, the frequency of a specific user's actual profile, namely q, will be given as the grouped count by categories of their check-ins. Finally, once the real profile of each user has been obtained, we discard those that do not present the desired features to evaluate our generalization technique (e.g., users presenting activity in a single category), and the datasets are ready for our experiments.

C. Results

In this section, we examine the extent to which our technique contributes to privacy preservation, in relation to the state-of-the-art perturbation techniques optimal suppression and optimal forgery. For this purpose, we analyze the effect of such techniques when the whole population of users enhances their privacy by using a common data perturbation rate, a simplified measure of loss in data utility that we denote by γ, σ and ρ,



(a) Foursquare



(b) Brightkite

Fig. 9: Geographic maps from experimental datasets over New York City area.

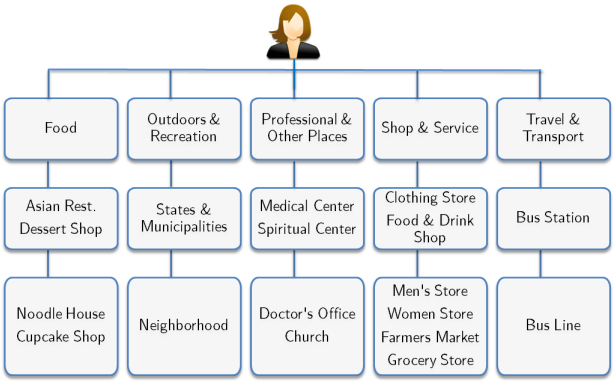
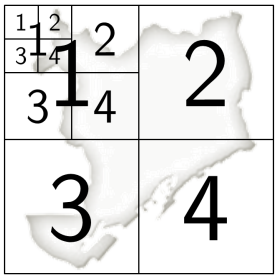


Fig. 10: Example of semantical hierarchy from Foursquare.

respectively, and show the privacy protection achieved by these users in terms of percentile curves (10th, 50th and 90th) of relative privacy gain. Recall that, in the architecture proposed in Sec. IV, a user specifies a generalization rate indicating the fraction of data he/she is disposed to generalize. Accordingly, we extend this idea to a common perturbation rate, where the user indicates the fraction of data that he/she is willing to perturb.

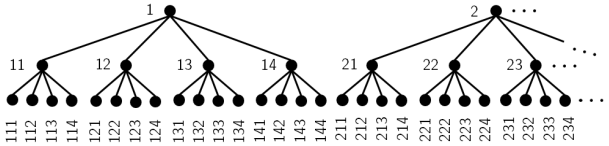
Under these assumptions, the first set of experiments consists in applying our optimal generalization technique, as well as the optimal suppression and forgery techniques, to all preselected users of both datasets, with spatial and semantic



(a) Recursive partitions

111	112	121	122	211	212	221	222
113	114	123	124	213	214	223	224
131	132	141	142	231	232	241	242
133	134	143	144	233	234	243	244
311	312	321	322	411	412	421	422
313	314	323	324	413	414	423	424
331	332	341	342	431	432	441	442
333	334	343	344	433	434	443	444

(b) Labeled partitions



(c) Hierarchical partitions

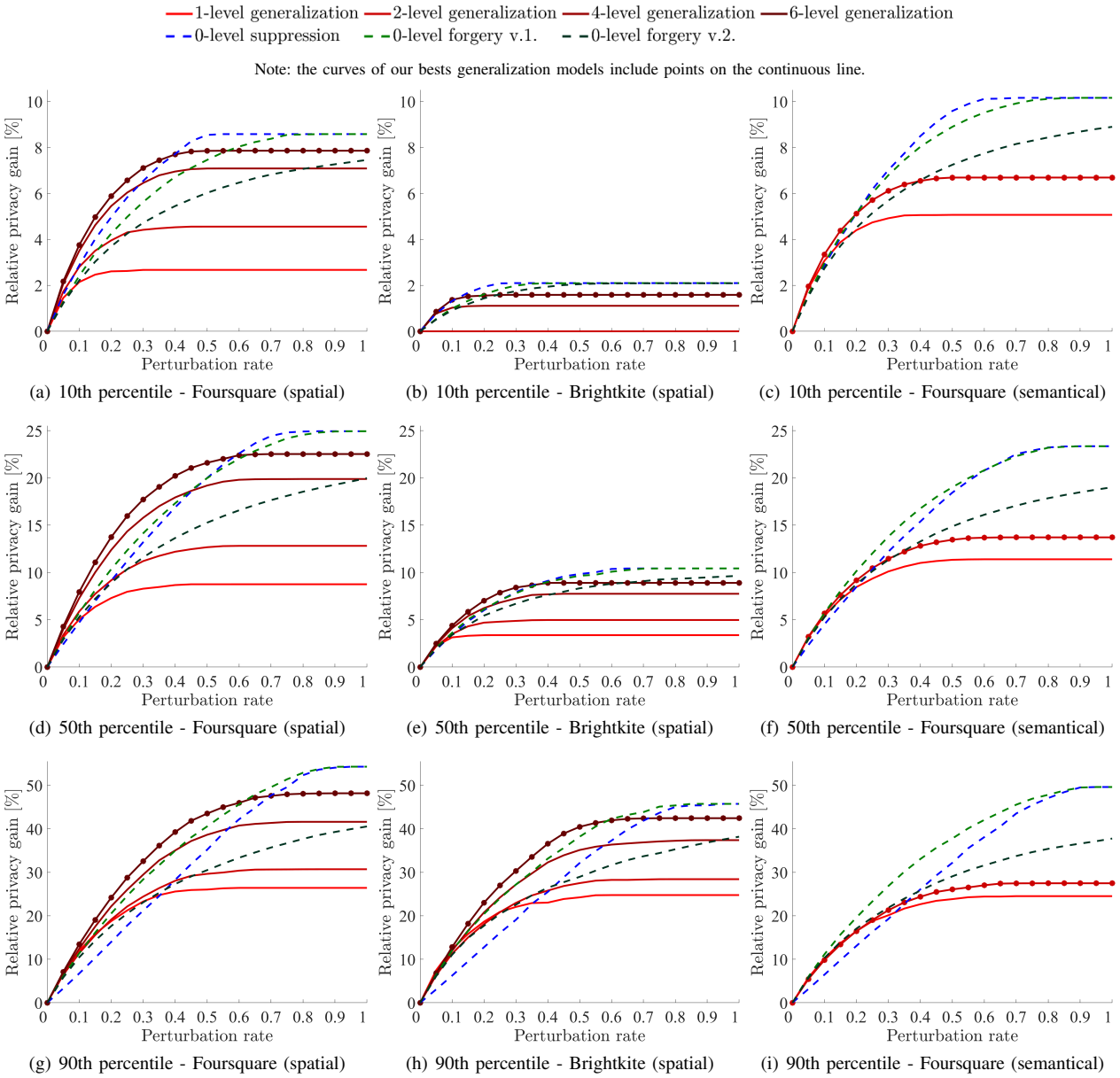
Fig. 11: Conceptual recursive partitions from 3-level hierarchical region division in quadrants.

(only in the case of Foursquare) type categorization of their interests, in accordance with Sec. VI-B.

For both data sets, Figs. 12 show the 10%, 50% and 90% percentile curves of relative privacy gain for generalization, using different hierarchy levels ($d = 1, 2, 4$ and 6 in the spatial case, $d = 1$ and 2, in the semantic case), for optimal suppression [15], and for optimal forgery (v.1 [12]), seven cases in total per dataset; we discarded forgery v.2. because of its low capacity for convergence within the fixed range of perturbation rates.

From those three figures, we can highlight that, in all cases, generalization with spatial and semantic hierarchical categorization of different levels can provide gains in relative privacy with *lower costs* (i.e., lower perturbation rates are needed to obtain the same privacy level) than the non-hierarchical techniques of suppression and forgery v.1., which both have a similar behavior. It should be noted, however, that, by definition, suppression and forgery always offer greater privacy in absolute terms ($\ln(n)$ is its maximum attainable value), in accordance with Corollary 2; and this is a characteristic also of the generalization technique which is maintained as the adherence of the number of users applying it increases. In addition, we can see that high-level generalization ($d = 6$) is the most competitive strategy compared to non-hierarchical techniques, although it is obvious from the results that the higher the level, the higher the cost.

Regarding the type of hierarchical categorization of our generalization technique, spatial or semantic, we observe that for the levels tested here ($d = 1$ and $d = 2$) the spatial hierarchy provides better relative gains compared to the semantic one. It should be noted at this point that the essential difference between the two is due to the size of the category groups, four fixed categories (quadrants) per group in the spatial type versus a variable number of categories per group in the semantic type.



(a) 10th percentile - Foursquare (spatial) (b) 10th percentile - Brightkite (spatial) (c) 10th percentile - Foursquare (semantical)

(d) 50th percentile - Foursquare (spatial) (e) 50th percentile - Brightkite (spatial) (f) 50th percentile - Foursquare (semantical)

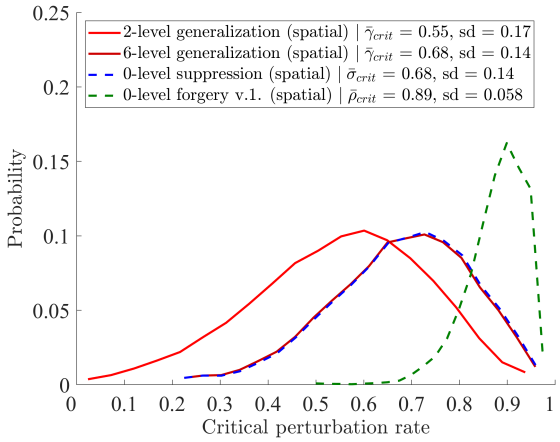
(g) 90th percentile - Foursquare (spatial) (h) 90th percentile - Brightkite (spatial) (i) 90th percentile - Foursquare (semantical)

Fig. 12: First experiment. 10, 50 and 90th percentile curves of relative privacy gain in the Foursquare (spatial and semantical) and Brightkite (spatial) datasets. Generalization is evaluated against the state-of-the-art techniques of optimal suppression and optimal forgery.

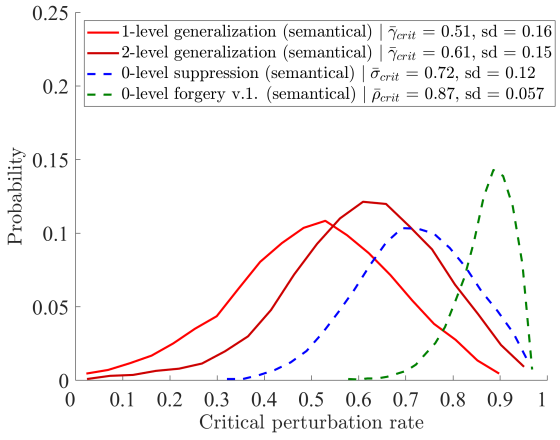
The second set of experiments focuses on the critical data perturbation rates of the analyzed and compared techniques, that is, our technique versus the state-of-the-art proposals of suppression and forgery. For this purpose, we analyze these rates collected in the first set of experiments in terms of their probability distribution. Fig. 13 show the results of these measurements in the form of a histograms for the spatial and semantical experiments, respectively. The first issue that seems remarkable to us is that the distribution of the critical generalization rate, called γ, is distributed in appearance as a normal variable, in the same sense as the critical rate of suppression, called σ, and versus the right-skewed form of the critical forgery rate, called ρ. In addition, the central tendency of our technique is generally more favorable, that is, lower than for non-hierarchical techniques, especially against forgery v.1. These results confirm our first set of experiments and hint at favorable and robust statistical properties of the critical perturbation rate of our technique, especially when establishing the most affordable uniform profiles.

Compared to suppression and forgery, we would like to highlight, in view of our experimental analysis on Foursquare and Brigthkite and the numerical example of Sec. V-B, that both techniques are more sensitive to the specific user profile than generalization in terms of the critical perturbation rate. Note that forgery tends to rates very close to 1, while in the theoretical results this circumstance occurs with suppression.

In essence, these experimental results confirm our hypotheses that the generalization of user data into categories is a technique that can protect user privacy, always at the cost of the utility of the system but in a quantifiable and therefore adjustable way, as well as being competitive against other non-hierarchical techniques. Finally, the results shown in this section illustrate how our mechanism perturbs the profile of the user observed from the outside and how this perturbation allows users to protect their privacy to a certain extent.



(a) Foursquare (spatial)



(b) Foursquare (semantical)

Fig. 13: Second experiment. Density plots of critical perturbation rates from Foursquare (spatial and semantical) dataset. Generalization rate is evaluated against state-of-art techniques of optimal supression and optimal forgery v.1.

In closing, the results shown in this section illustrate how our mechanism perturbs the user profile observed from the outside and how this perturbation enables users to protect their privacy to a certain degree.

D. Discussion

The experiments carried out have served to test our proposed generalization mechanism in a demanding and rigorous way. This requirement is combined with the robustness of the mathematical framework in which we model and theoretically analyze the optimized version of this mechanism. Although we do not offer a closed solution of the generalization strategy g^* for any generalization rate γ, we provide our theoretical analysis with some important characteristics of the mechanism, such as the critical generalization rate γ_{crit} or the upper bound of the value of privacy for that critical rate times the logarithm of the profile dimension n. And we can add that in this work we have gone one step further by evaluating our proposal against other mechanisms that are part of the state-of-the-art of privacy data-perturbative techniques. In this sense, we consider that these premises allow us to offer a positive experience in terms of the validity and accuracy of the results obtained.

The behavior of our mechanism is partly expected. Although we did not prove it analytically, we intuit that our privacy–generalization function $\mathcal{P}(\gamma)$ is nondecreasing and quasiconcave, and we confirm this assumption by demonstrating that there is a critical generalization rate γ_{crit}, as we can also observe in the graphical representations. Or what is the same, beyond a certain percentage of generalized data, privacy does not increase. However, we are struck by the pleasant simplicity of the explicit value of the critical generalization rate taking into account the extensive nature of the hierarchy used in the problem.

We find it remarkable from the spatial experiment on the relative gain of privacy that, with a good level of hierarchy, $d = 6$, in this case, the privacy provided for 90% of the users is so competitive in terms of privacy and at the same time so profitable in terms of data utility. We should expect the same effect in the semantic case although we have not been able to evaluate a taxonomy with such a high hierarchical level. This trend is later confirmed with the experiment on critical perturbation rates. And as we have explained, the reason lies both in the way we perturb the real user profile q and in its distribution.

Finally, we want to signify that our results present two relevant contributions to the state-of-the-art of data-perturbation mechanisms in the context of PISs. First, we contribute a new strong privacy mechanism in this area, especially in the case that the data is subject to some type of hierarchical taxonomy, although we have already seen that we can also apply our mechanism spatially. And secondly, we demonstrate that if such a hierarchy exists, generalization can be the most profitable alternative in terms of utility compared to pre-existing mechanisms. In short, we want to expand the range of possibilities so that PISs designers/engineers can evaluate the performance of their systems and adopt the ideal mechanism based on their needs. It is worth remembering in any case as an extra reasoning, that we offer a mechanism that seeks to protect the privacy of each user individually, compared to mechanisms that, for example, seek average privacy for groups of users.

VII. Conclusions and Future Work

We have proposed data generalization as a simple data minimization strategy against non-fully trusted personalized information systems. Our proposal does not require users to trust an external entity nor the network operator, and can be used in combination with other privacy mechanisms such

as traditional anonymous-communication systems. However, generalization comes at the cost of some processing overhead but more importantly at the expense of semantic loss incurred by generalizing data. In other words, generalization data poses an inherent trade-off between privacy and data utility.

Our main contribution is, precisely, a systematic, mathematical approach to the problem of optimal data generalization. We have measured user privacy as the entropy of the user's apparent data distribution, after the generalization of data. Subsequently, we have formulated an optimization problem modeling the privacy–utility trade-off.

In our mathematical model, we have represented user data as r.v.'s taking on values on a common finite alphabet of categories. This has allowed us to describe user profiles as PMFs, a representation that is frequently used in information systems. The proposed model, however, is restricted to relative frequencies and hence does not deal with differences in the absolute frequencies. Besides, we have assumed, on the one hand, a simple attacker model where the information system (the adversary) is not able to estimate a particular user's generalization rate, and on the other, that only a small number of users adhere to this technique.

Our theoretical analysis has proven that, under the positivity assumption (4), there exists a critical generalization rate γ_{crit}, beyond which critical privacy is achievable. Specifically, we have shown that this γ_{crit} only depends on the number of bottom-level categories by top-level category in a predefined hierarchy and the first component of actual profile q in the mentioned category groups.

We have proposed a system architecture that implements data generalization in practice, and investigated its application into two real-world datasets. Specifically, we have assessed experimentally the extent to which our solution may help users protect their privacy and evaluated it versus optimal suppression and optimal forgery, two state-of-the-art perturbation techniques in information systems, showing that generalization can outperform these two by incurring less information loss for a same privacy level.

ACKNOWLEDGMENTS

Javier Parra-Arnau is the recipient of a "Ramón y Cajal" fellowship (ref. RYC2021-034256-I) funded by the Spanish Ministry of Science and Innovation and the European Union – "NextGenerationEU"/PRTR (Plan de Recuperación, Transformación y Resiliencia). This work was also supported by the Spanish Government under the project "Enhancing Communication Protocols with Machine Learning while Protecting Sensitive Data (COMPROMISE)" PID2020-113795RB-C31, funded by MCIN/AEI/10.13039/501100011033; the project "Anonymization technology for AI-based analytics of mobility data (MOBILYTICS)" (TED2021-129782B-I00), funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR; and funded by the Generalitat de Catalunya, under AGAUR grant "2021 SGR 01413".

REFERENCES

[1] I. T. Union, "Measuring digital development: Facts and figures 2020," 2020.

[2] K. Schwab, "The fourth industrial revolution: what it means, how to respond," *Foreign Affairs*, vol. 12, pp. 2015–17, 2015.

[3] K. Boeckl, K. Boeckl, M. Fagan, W. Fisher, N. Lefkovitz, K. N. Megas, E. Nadeau, D. G. O'Rourke, B. Piccarreta, and K. Scarfone, *Considerations for managing Internet of Things (IoT) cybersecurity and privacy risks*. US Department of Commerce, National Institute of Standards and Technology, 2019.

[4] N. Kaaniche, M. Laurent, and S. Belguith, "Privacy enhancing technologies for solving the privacy-personalization paradox: Taxonomy and survey," *Journal of Network and Computer Applications*, p. 102807, 2020.

[5] G. Danezis, "Introduction to privacy technology," *Katholieke University Leuven, COSIC: Leuven, Belgium*, 2007.

[6] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen, "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," *Requirements Engineering*, vol. 16, no. 1, pp. 3–32, 2011.

[7] A. Hundepool, A. V. de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. de Wolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing, "μ-ARGUS version 3.2 software and user's manual," Voorburg, Netherlands, 2003. [Online]. Available: http://neon.vb.cbs.nl/casc

[8] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-enhancing personalized Web search," in *Proc. Int. WWW Conf.* ACM, 2007, pp. 591–600.

[9] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy preserving targeted advertising," in *Proceedings Network and Distributed System Symposium*, 2010.

[10] ——, "Adnostic: Privacy preserving targeted advertising," in *Proc. Symp. Netw. Distrib. Syst. Secur. (SNDSS)*, Feb. 2010, pp. 1–21.

[11] M. Fredrikson and B. Livshits, "RePriv: Re-envisioning in-browser privacy," in *Proc. IEEE Symp. Secur., Priv. (SP)*, May 2011, pp. 131–146.

[12] D. Rebollo-Monedero and J. Forné, "Optimized query forgery for private information retrieval," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4631–4642, 2010.

[13] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A privacy-preserving architecture for the semantic web based on tag suppression," in *International Conference on Trust, Privacy and Security in Digital Business*. Springer, 2010, pp. 58–68.

[14] ——, "A privacy-protecting architecture for collaborative filtering via forgery and suppression of ratings," in *Data Privacy Management and Autonomous Spontaneus Security*. Springer, 2011, pp. 42–57.

[15] J. Parra-Arnau, D. Rebollo-Monedero, J. Forné, J. L. Muñoz, and O. Esparza, "Optimal tag suppression for privacy protection in the semantic web," *Data & Knowledge Engineering*, vol. 81, pp. 46–66, 2012.

[16] J. Parra-Arnau, A. Perego, E. Ferrari, J. Forne, and D. Rebollo-Monedero, "Privacy-preserving enhanced collaborative tagging," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 180–193, 2012.

[17] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Optimal forgery and suppression of ratings for privacy enhancement in recommendation systems," *Entropy*, vol. 16, no. 3, pp. 1586–1631, 2014.

[18] ——, "Measuring the privacy of user profiles in personalized information systems," *Future Generation Computer Systems*, vol. 33, pp. 53–63, 2014.

[19] A. Rodriguez-Carrion, D. Rebollo-Monedero, J. Forne, C. Campo, C. Garcia-Rubio, J. Parra-Arnau, and S. K. Das, "Entropy-based privacy against profiling of user mobility," *Entropy*, vol. 17, no. 6, pp. 3913–3946, 2015. [Online]. Available: https://www.mdpi.com/1099-4300/17/6/3913

[20] J. Estrada-Jimenez, J. Parra-Arnau, A. Rodriguez-Hoyos, and J. Forne, "On the regulation of personal data distribution in online advertising platforms," *Eng. Appl. Artif. Intell.*, vol. 82, pp. 13–29, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197619300636

[21] J. Parra-Arnau, "Pay-per-tracking: A collaborative masking model for web browsing," *Information Sciences*, vol. 385-386, pp. 96–124, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025516322587

[22] ——, "Optimized, direct sale of privacy in personal data marketplaces," *Information Sciences*, vol. 424, pp. 354–384, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025517310022
[23] J. Parra-Arnau, A. Perego, E. Ferrari, J. Forne, and D. Rebollo-Monedero, "Privacy-preserving enhanced collaborative tagging," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 180–193, 2014.
[24] I. Ullah and A. Binbusayyis, "Joint optimization of privacy and cost of in-app mobile user profiling and targeted ads," *IEEE Access*, vol. 10, pp. 38 664–38 683, 2022.
[25] L. A. Leiva, I. Arapakis, and C. Iordanou, "My mouse, my rules: Privacy issues of behavioral user profiling via mouse tracking," in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 2021, pp. 51–61.
[26] L. Hernández-Álvarez, J. M. de Fuentes, L. González-Manzano, and L. Hernández Encinas, "Privacy-preserving sensor-based continuous authentication and user profiling: a review," *Sensors*, vol. 21, no. 1, p. 92, 2020.
[27] M. Mamun, M. Al-Digeil, and S. S. Ahmed, "Profiling online users: Emerging approaches and challenges," *Securing Social Networks in Cyberspace*, pp. 221–240, 2021.
[28] Shakil, M. Arif, S. S. Sohail, M. T. Alam, S. Ubaid, M. T. Nafis, and G. Wang, "Towards a two-tier architecture for privacy-enabled recommender systems (pers)," in *Inernational Conference on Ubiquitous Security*. Springer, 2021, pp. 268–278.
[29] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Privacy-enhancing technologies and metrics in personalized information systems," in *Advanced Research in Data Privacy*. Springer, 2015, pp. 423–442.
[30] Y. Elovici, B. Shapira, and A. Maschiach, "A new privacy model for hiding group interests while accessing the web," in *Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society*, 2002, pp. 63–70.
[31] ——, "A new privacy model for web surfing," in *International Workshop on Next Generation Information Technologies and Systems*. Springer, 2002, pp. 45–57.
[32] Y. Elovici, C. Glezer, and B. Shapira, "Enhancing customer privacy while searching for products and services on the world wide web," *Internet Research*, 2005.
[33] Y. Elovici, B. Shapira, and A. Meshiach, "Cluster-analysis attack against a private web solution (praw)," *Online Information Review*, 2006.
[34] S. Ye, F. Wu, R. Pandey, and H. Chen, "Noise injection for search privacy protection," in *2009 International Conference on Computational Science and Engineering*, vol. 3. IEEE, 2009, pp. 1–8.
[35] D. Howe and H. Nissenbaum, "Lessons from the identity trail: Privacy, anonymity and identity in a networked society. ny: Oxford univ. press, 2009, ch. trackmenot: Resisting surveillance in web search," *http://mrl.nyu.edu/ dhowe/trackmenot*, pp. 417–436, 2009.
[36] R. Chow and P. Golle, "Faking contextual data for fun, profit, and privacy," in *Proceedings of the 8th ACM workshop on Privacy in the electronic society*, 2009, pp. 105–108.
[37] J. Domingo-Ferrer, A. Solanas, and J. Castellà-Roca, "h(k)-private information retrieval from privacy-uncooperative queryable databases," *Online Information Review*, 2009.
[38] E. Balsa, C. Troncoso, and C. Diaz, "Ob-pws: Obfuscation-based private web search," in *2012 IEEE Symposium on Security and Privacy*. IEEE, 2012, pp. 491–505.
[39] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[40] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-enhancing personalized web search," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 591–600.
[41] M. Fredrikson and B. Livshits, "Repriv: Re-envisioning in-browser privacy," in *Proc. IEEE Symp. Security, Privacy (SP)(May 2011)*, 2010.
[42] J. Parra-Arnau, J. P. Achara, and C. Castelluccia, "Myadchoices: Bringing transparency and control to online advertising," *ACM Transactions on the Web (TWEB)*, vol. 11, no. 1, pp. 1–47, 2017.
[43] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.
[44] J. P. Arnau, "Privacy protection of user profiles in personalized information systems," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2014.
[45] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User profiles for personalized information access," *The adaptive web*, pp. 54–89, 2007.
[46] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
[47] A. Viejo, D. Sánchez, and J. Castella-Roca, "Using profiling techniques to protect the user's privacy in twitter," in *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 2012, pp. 161–172.
[48] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.
[49] ——, "Cvx: Matlab software for disciplined convex programming, version 2.1," 2014.
[50] M. ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.*, 2019. [Online]. Available: http://docs.mosek.com/9.0/toolbox/index.html
[51] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 1, pp. 129–142, 2014.
[52] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1082–1090.
[53] S. Oya, C. Troncoso, and F. Pérez-González, "Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1959–1972.