A Privacy-Protecting Architecture for Recommendation Systems via the Suppression of Ratings

Javier Parra-Arnau, David Rebollo-Monedero and Jordi Forné
Department of Telematics Engineering
Universitat Politècnica de Catalunya
C. Jordi Girona 1-3, E-08034 Barcelona, Spain
{javier.parra,david.rebollo,jforne}@entel.upc.edu

Abstract

Recommendation systems are information-filtering systems that help users deal with information overload. Unfortunately, current recommendation systems prompt serious privacy concerns. In this work, we propose an architecture that enables users to enhance their privacy in those systems that profile users on the basis of the items rated. Our approach capitalizes on a conceptually-simple perturbative technique, namely the suppression of ratings. In our scenario, users rate those items they have an opinion on. However, in order to avoid being accurately profiled, they may want to refrain from rating certain items. Consequently, this technique protects user privacy to a certain extent, but at the cost of a degradation in the accuracy of the recommendation.

We measure privacy risk as the Kullback-Leibler divergence between the user's and the population's rating distribution, a privacy criterion that we proposed in previous work. The justification of such a criterion is our second contribution. Concretely, we thoroughly interpret it by elaborating on the intimate connection between the celebrated method of entropy maximization and the use of entropies and divergences as measures of privacy. The ultimate purpose of this justification is to attempt to bridge the gap between the privacy and the information-theoretic communities by substantially adapting some technicalities of our original work to reach a wider audience, not intimately familiar with information theory and the method of types. Lastly, we present a formulation of the optimal trade-off between privacy and suppression rate, what allows us to formally specify one of the functional blocks of the proposed architecture.

1 Introduction

From the advent of the Internet and the World Wide Web (WWW), the amount of information available to users has grown exponentially. Today, due to this information overload, users feel they have to separate the wheat from the chaff. Recommendation systems are a type of information-filtering systems that assist users in this task by suggesting information items they may be interested in. Examples of these systems include recommending books, music, and other products at Amazon.com [1], movies by MovieLens [2] and Netflix [3], and news at Digg [4].

One of the most popular forms of interaction in recommendation systems is that users communicate their preferences by rating items. This is the case of Movielens, where users

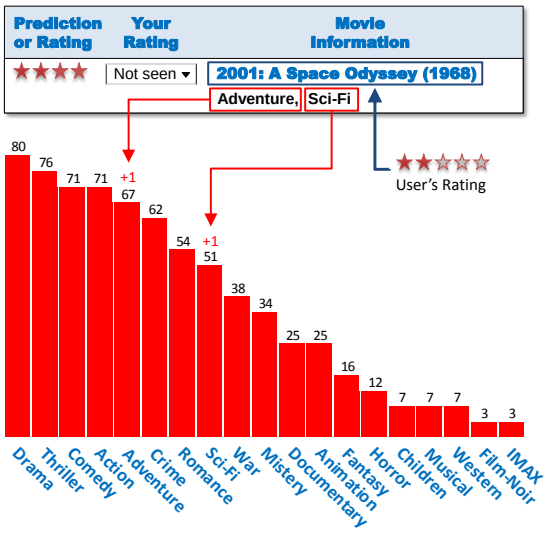

Figure 1. The profile of a user is modeled in Movielens as a histogram of absolute frequencies of ratings within a set of movie genres (bottom). Based on this profile, the recommender predicts the rating that the user would probably give to a movie (top). After having watched the movie, the user rates it and their profile is updated.

assign *ratings* to movies they have already watched. Other strategies to capture users' interests include asking them to sort a number of items by order of predilection, or suggesting that they mark the items they like. On the other hand, recommendation systems may collect data from users without requiring them to explicitly convey their interests [5]. Such practices include observing the items clicked by users in an online store, analyzing the time it takes users to examine an item, or simply keeping a record of the purchased items.

The prolonged collection of these data allows the system to extract an accurate snapshot of user interests or *user profiles*. Once this information has been captured, the recommendation system applies an algorithm that returns a prediction of users' interests for those items they have not yet considered. For example, Movielens and Digg apply collaborative-filtering algorithms [6, 7] to predict the rating that a user would give to a movie and to create a personalized list of recommended news, respectively. Fig. 1 illustrates the case of Movielens and provides an example of user profile.

Despite the many advantages recommendation systems are bringing to users, the information collected, processed and stored by these systems prompts serious privacy concerns. One of the main privacy risks perceived by users is that of a computer "figuring things out" about them [8]. Namely, many users are worried about the idea that their profiles may reveal sensitive information such as health-related issues, political preferences, salary or religion. On the other hand, other users are concerned that the system's predictions may be totally erroneous and be later used to defame them. The latter situation is illustrated in [9], where the accuracy of the predictions provided by TiVo digital video recorder and Amazon is questioned. Specifically, the author describes several real cases in which the recommender makes dubious, and in some cases aberrant, inferences about users' sexual preferences. Lastly, other privacy risks embrace unsolicited marketing, information leaked to other users of the same computer, court subpoenas, and government surveillance [8].

Consequently, it is not surprising that some users are reticent to disclose their interests. Actually, a report [10] finds that 95% of the respondents refused, at some point, to provide personal information when requested by a Web site. In a nutshell, this just reinforces the fact that refusing to give private information may be considered as a strategy accepted by users concerned with their privacy.

1.1 Contribution and Plan of this Paper

In this work, we tackle the problem of protecting user profiles in recommendation systems. With this purpose, we present an architecture that enables users to enhance their privacy in those systems where they are profiled on the basis of their ratings. Our approach relies upon a conceptually-simple mechanism, namely the suppression of ratings. In our scenario, users rate those items they have an opinion on. However, in order to avoid being accurately profiled, they may wish to refrain from rating certain items. Therefore, this approach protects user privacy to a certain extent, without having to trust the recommendation system or the network operator, but at the cost a loss in utility, a degradation in the accuracy of the prediction.

We measure user privacy as the Kullback-Leibler (KL) divergence between the user's and the population's rating distribution, a criterion that we presented in previous work [11]. Our second contribution is precisely the interpretation of this privacy metric, which contemplates the entropy of the user's item distribution as a particular case. In this work, we thoroughly justify this measure, by elaborating on the intimate connection between the celebrated method of entropy maximization and the use of entropies and divergences as measures of privacy. This justification also attempts to bridge the gap between the privacy and the information-theoretic communities by substantially adapting some technicalities of our original work to reach a wider audience, not intimately familiar with information theory and the method of types.

In addition, we present an information-theoretic, mathematical formulation of the trade-off between privacy and suppression rate. Our formulation results in a convex optimization problem for which there exist efficient numerical methods to solve it. Last but not least, we would like to stress that our approach could benefit from the combination with other alternatives in the literature.

Sec. 2 reviews some relevant approaches aimed at preserving user privacy in recommendation systems. Sec. 4 describes a privacy-protecting architecture based on the suppression of ratings. In addition, this section presents the model of user profile assumed, the adversarial model and our privacy measure. It is not until Sec. 5 where we shall carefully justify this privacy metric. Later in Sec. 6 we introduce a formulation of the trade-off between privacy and suppression rate. Conclusions are drawn in Sec. 7.

2 State of the Art

In this section, first we overview some of the most prominent privacy mechanisms in the motivating scenario of this work, namely recommendation systems; and secondly, we touch upon the most popular privacy metrics.

2.1 Privacy-Enhancing Mechanisms

Numerous approaches have been proposed to protect user privacy in the context of recommendation systems. These approaches basically suggest three main strategies: perturbing the information provided by users, using cryptographic techniques, and distributing the information stored by recommenders.

In the case of perturbative methods for recommendation systems, [12] proposes that users add random values to their ratings and then submit these perturbed ratings to the recommender. After receiving these ratings, the system executes an algorithm and sends the users some information that allows them to compute the prediction. When the number of participating users is sufficiently large, the authors find that user privacy is protected to a certain extent and the system reaches a decent level of accuracy. However, even though a user disguises all their ratings, it is evident that the items themselves may uncover sensitive information. In other words, the simple fact of showing interest in a certain item may be more revealing than the ratings assigned to that item. For instance, a user rating a book called "How to Overcome Depression" indicates a clear interest in depression, regardless of the score assigned to this book. Apart from this critique, other works [13, 14] stress that the use of *randomized* data distortion techniques might not able to preserve privacy.

In line with this work, [15] applies the same perturbative technique to CF algorithms based on singular-value decomposition (SVD). More specifically, the authors focus on the impact that their technique has on privacy. For this purpose, they use the privacy metric proposed by [16], which is essentially equivalent to *differential entropy*, and conduct some experiments with data sets from Movielens and Jester [17]. The results show the trade-off curve between accuracy in recommendations and privacy. In particular, they measure accuracy as the mean absolute error between the predicted values from the original ratings and the predictions obtained from the perturbed ratings.

At this point, we would like to remark that the use of perturbative techniques is by no means new in other application scenarios such as private information retrieval (PIR). In this scenario, users send general-purpose queries to an information service provider. An example would be a user sending the query "What was George Orwell's real name?". A perturbative approach to protect user profiles in this context consists in combining genuine with false queries. In this sense, [11] proposes a *non-randomized* method for query forgery and investigates the trade-off between privacy and the additional traffic overhead.

Regarding the use of cryptographic techniques, [18, 19] propose a method that enables a community of users to calculate a public aggregate of their profiles without revealing them on an individual basis. In particular, the authors use a homomorphic encryption scheme and a peer-to-peer (P2P) communication protocol for the recommender to perform this calculation. Once the aggregated profile is computed, the system sends it to users, who finally use local computation to obtain personalized recommendations. This proposal prevents the system or any external attacker from ascertaining the individual user profiles. However, its main handicap is assuming that an acceptable number of users is online and willing to participate in the protocol. In line with this, [20] uses a variant of Pailliers' homomorphic cryptosystem which improves the efficiency in the communication protocol. Another solution [21] presents an algorithm aimed at providing more efficiency by using the scalar product protocol.

In order to mitigate the potential privacy risks derived from the fact that users' private information is kept in a single repository, some approaches suggest that this information be

stored in a distributed way. This is the case of [22], which presents a CF algorithm called PocketLens, specifically designed to be deployed to a P2P scenario. The algorithm in question enables users to decide which private information should exchange with other users of the P2P community. In addition, the authors provide several architectures for the problem of locating neighbors. Another alternative assumes a pure decentralized P2P scenario and proposes the use of several perturbative strategies [23]. In essence, this scheme could be regarded as a combination of the approaches in [22] and [12]. Namely, the mentioned scheme recommends replacing the actual ratings by fixed, predefined values, by uniformly distributed random values, and by a bell-curve distribution imitating the distribution of the population's ratings.

2.2 Privacy Criteria

In this section we give a broad overview of privacy criteria originally intended for statistical disclosure control (SDC), but in fact applicable to the domain of recommendation systems, the motivating application of our work. In database privacy, a *microdata set* is defined as a database table whose records carry information concerning individual respondents. Specifically, this set contains key attributes, that is, attributes that, in combination, may be linked with external information to reidentify the respondents to whom the records in the microdata set refer. Examples include job, address, age and gender, height and weight. In addition, the data set contains confidential attributes with sensitive information on the respondent, such as health, salary and religion.

A common approach in SDC is microaggregation, which consists in clustering the data set into groups of records with similar tuples of key attributes values, and replacing these tuples in every record within each group by a representative group tuple. One of the most popular privacy criteria in database anonymization is k-anonymity [24], which can be achieved through the aforementioned microaggregation procedure. This criterion requires that each combination of key attribute values be shared by at least k records in the microdata set. However, the problem of k-anonymity, and of enhancements [25–28] such as l-diversity, is their vulnerability against skewness and similarity attacks [29]. In order to overcome these deficiencies, yet another privacy criterion was considered in [30]: a dataset is said to satisfy t-closeness if for each group of records sharing a combination of key attributes, a certain measure of divergence between the within-group distribution of confidential attributes and the distribution of those attributes for the entire dataset does not exceed a threshold t. An average-case version of the worst-case t-closeness criterion, using the Kullback-Leibler divergence as a measure of discrepancy, turns out to be equivalent to a mutual information, and lend itself to a generalization of Shannon's rate-distortion problem [31, 32].

A simpler information-theoretic privacy criterion, not directly evolved from k-anonymity, consists in measuring the degree of anonymity observable by an attacker as the entropy of the probability distribution of possible senders of a given message [33, 34]. A generalization and justification of such criterion, along with its applicability to PIR, are provided in [11,35].

3 Statistical and Information-Theoretic Preliminaries

This section establishes notational aspects, and, in order to make our presentation suited to a wider audience, recalls key information-theoretic concepts assumed to be known in the remainder of the paper, specially in Sec. 5 where we justify our privacy metric.

The measurable space in which a *random variable* (r.v.) takes on values will be called an *alphabet*, which, with a mild loss of generality, we shall always assume to be finite. We shall follow the convention of using uppercase letters for r.v.'s, and lowercase letters for particular values they take on. The *probability mass function* (PMF) p of an r.v. X is essentially a *relative histogram* across the possible values determined by its alphabet. Informally, we shall occasionally refer to the function p by its value $p(x)$. The *expectation* of an r.v. X will be written as $\mathrm{E}\,X$, concisely denoting $\sum_x x\,p(x)$, where the sum is taken across all values of x in its alphabet.

We adopt the same notation for information-theoretic quantities used in [36]. Concordantly, the symbol H will denote entropy and D relative entropy or KL divergence. We briefly recall those concepts for the reader not intimately familiar with information theory. All logarithms are taken to base 2. The *entropy* $\mathrm{H}(p)$ of a discrete r.v. X with probability distribution p is a measure of its uncertainty, defined as

$$\mathrm{H}(X) = -\,\mathrm{E}\,\log p(X) = -\sum_x p(x)\log p(x).$$

Given two probability distributions $p(x)$ and $q(x)$ over the same alphabet, the *KL divergence* or *relative entropy* $\mathrm{D}(p\,\|\,q)$ is defined as

$$\mathrm{D}(p\,\|\,q) = \mathrm{E}_p\,\log\frac{p(X)}{q(X)} = \sum_x p(x)\log\frac{p(x)}{q(x)}.$$

The KL divergence is often referred to as *relative entropy*, as it may be regarded as a generalization of entropy of a distribution, relative to another. Conversely, entropy is a special case of KL divergence, as for a uniform distribution u on a finite alphabet of cardinality n,

$$\mathrm{D}(p\,\|u) = \log n - \mathrm{H}(p). \tag{1}$$

Although the KL divergence is not a distance in the mathematical sense of the term, because it is neither symmetric nor satisfies the triangle inequality, it does provide a measure of discrepancy between distributions, in the sense that $\mathrm{D}(p\,\|\,q) \geq 0$, with equality if, and only if, $p = q$. On account of this fact, relation (1) between entropy and KL divergence implies that $\mathrm{H}(p) \leqslant \log n$, with equality if, and only if, $p = u$. Simply put, *entropy maximization* is a special case of *divergence minimization*, attained when the distribution taken as optimization variable is identical to the *reference distribution*, or as "close" as possible, should the optimization problem appear accompanied with *constraints* on the desired space of candidate distributions.

4 Privacy Protection in Recommendation Systems via the Suppressing of Ratings

In this section, we present our first contribution: an architecture for the protection of user profiles in recommendation systems. Particularly, we consider the case in which users' preferences are exclusively derived from the ratings they assign to items. Our approach is based on a perturbative technique, namely the suppression of ratings to items. In our scenario, users rate items according to their personal preferences. However, in order to avoid being accurately profiled, they may want to refrain from rating some of those items.

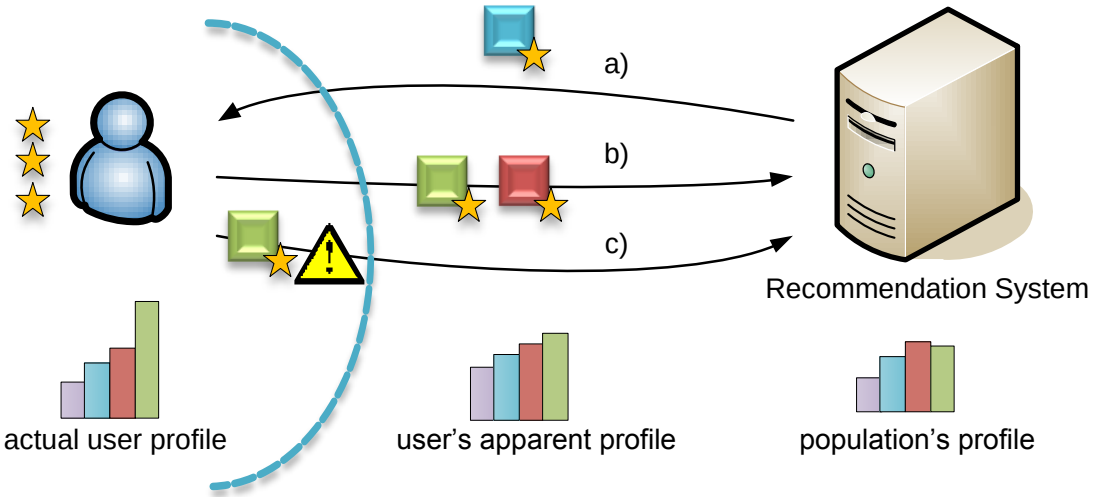

Figure 2. A user retrieves a particular item and the ratings submitted by the other users to it, from a recommendation system (a). Later, the user submits their own ratings to such recommender (b). Afterwards, the user receives a privacy alarm when trying to submit a new rating (c), because their actual profile deviates significantly from the population distribution of ratings.

We would like to stress that our approach could be integrated with other systems, like for example, with some of the approaches mentioned in Sec. 2, and those using pseudonyms [37, 38].

In the rest of this section, we provide further insight into our proposal. Concretely, we propose a mathematical model of user profiles in Sec. 4.1. Afterwards, Sec. 4.2 examines the assumed adversarial model. Next, our privacy criterion is presented in Sec. 4.3, but it is not until Sec. 5 when we shall thoroughly justify this metric. Lastly, we delve into our architecture and analyze each of its internal components in Sec. 4.4.

4.1 User Profile

We pointed out in Sec. 1 that Movielens uses histograms of absolute frequencies to show user profiles. Other systems such as Jinni and Last.fm represent this information by means of a tag cloud, which may be regarded as another kind of histogram. In this spirit, recent privacy-protecting approaches in the scenario of recommendation systems propose using histograms of absolute frequencies [39, 40].

According to all these examples, and as used in [11, 35, 41], we propose a tractable model of user profile as a PMF, that is, a histogram of relative frequencies of ratings within a predefined set of categories of interest. We would like to remark that, under this model, user profiles do not capture the particular scores given to items, but what we consider to be more sensitive: the categories these items belong to. This corresponds to the case of Movielens, which we illustrate in Fig. 1. In this example, a user assigns two stars to a movie, meaning that they consider it to be "fairly bad". However, the recommender updates their profile based only on the categories this movie belongs to.

Having assumed the above model, now we focus on how to estimate the profile of a user from their ratings. The reason is that our approach requires this information to help

users decide which items should be rated and which should not. Clearly, the easiest way to obtain a user profile is by asking the recommender. Movielens users, for instance, can do that. Unfortunately, in most recommendation systems users do not have access to this information. In order to cope with this, we suggest an alternative for extracting users' preferences from their rating activity.

We consider two possible cases for the information that a system shows about its items. The trivial case is when the recommender provides users with a categorization of all of its items. In this situation, it is straightforward to keep a histogram based on these categories. This is the case of Netflix or Movielens, where the genres of all movies are available to users. On the contrary, it may happen that this categorization is not at the disposal of users. This applies to Digg, where the only information that the recommender provides about news is the headline, the first lines of the news and the source of information. In systems like this, the categorization of items may be accomplished by exploring web pages with information about those items. Specifically, this process could be carried out by using the vector space model [42], as normally done in information retrieval, to represent these web pages as tuples containing their most representative terms. Namely, the term frequency-inverse document frequency (TF-IDF) could be applied to calculate the weights of each term appearing in a web page. Next, the most weighted terms of each web page could be combined in order to create a category and assign it to the item. After obtaining the categories associated with all the items rated by a user, their profile would be computed as a histogram across these categories.

4.2 Adversarial Model

In our scenario, we suppose users interact with recommendation systems that infer their preferences based only on their ratings. This supposition is reinforced by the tractability of the model considered and also by the fact that implicit mechanisms are often less accurate than explicit ratings [43].

Under this assumption, we consider an adversarial model in which users submitting their ratings are observed by a passive attacker who is able to ascertain which ratings are associated with which items. Concretely, this could be the case of the recommendation system itself or, in general, any privacy attacker able to crawl through this information.

Bearing in mind the model of user profile assumed in Sec. 4.1, after the rating of a sufficiently large number of items, the attacker can compute a histogram with the actual interests of a particular user. However, when this user adheres to the suppression of ratings, the attacker observes a perturbed version of this histogram, which makes it more difficult for the attacker to discover the user's actual preferences. We shall refer to this perturbed profile as the user's *apparent* profile. Last but not least, we suppose that the attacker is unaware of or ignores the fact that the user is adopting our strategy, thereby assuming that the apparent profile reflects genuine interests.

4.3 Privacy Metric

Any optimized mechanism aimed at protecting the privacy of users necessarily requires to evaluate the extent to which it is effective. In this work, just as in [11, 35], we use an information-theoretic quantity to emphasize that an attacker will have gained some

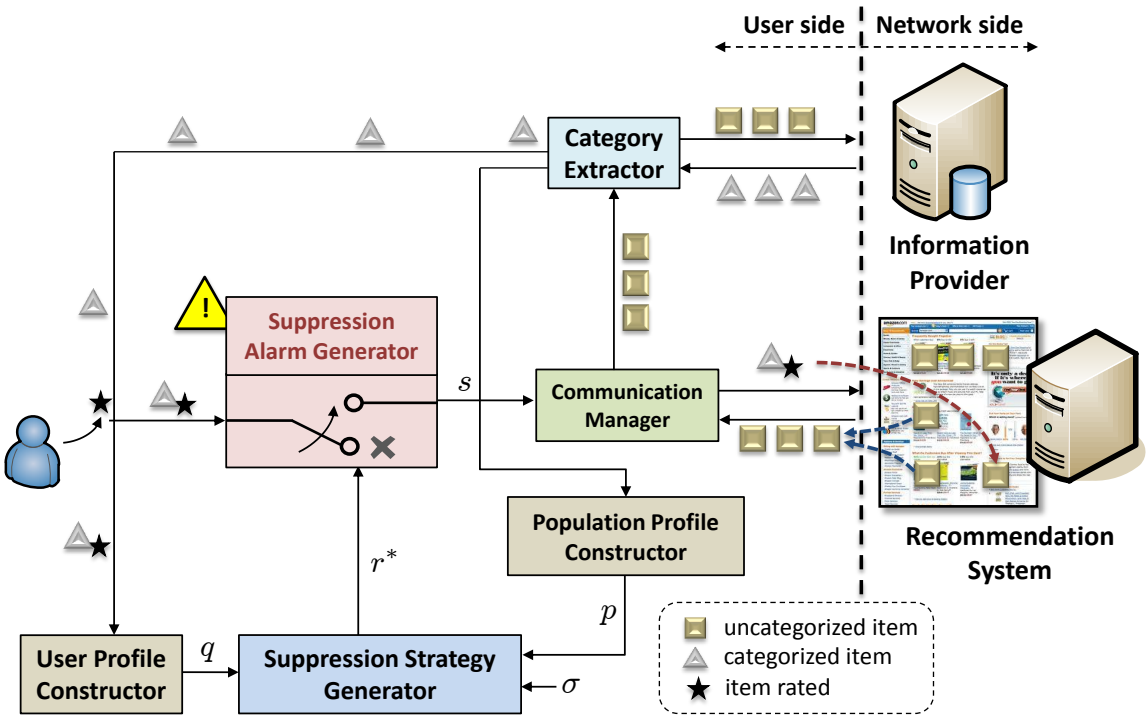

Figure 3. Block diagram of the proposed architecture.

information about a user whenever their preferences diverge from the average population interests.

Specifically, inspired by the privacy criteria proposed in [32], we consider the KL divergence [36], introduced in Sec. 3, which may be interpreted as a measure of discrepancy between probability distributions. Accordingly, we measure privacy risk as the KL divergence between the apparent profile s resulting from the elimination of certain ratings and the population distribution of ratings p, namely $\mathrm{D}(s \,\|\, p)$. The justification and interpretation of this measure of privacy is the purpose of Sec. 5. A formulation of the trade-off between privacy and suppression is presented later in Sec. 6.

4.4 Architecture

In this section, we describe an architecture that helps users decide which items should be rated and which should not, in order to hinder privacy attackers in their efforts to profile users' interests. Our architecture is conceived to be implemented by a software application running on the user's local machine. Fig. 3 shows the proposed architecture, which consists of a number of modules, each of them performing a specific task. Next, we provide a functional description of all of its modules and examine the details of a practical implementation.

Communication Manager. This module is in charge of interacting with the recommendation system. Specifically, it downloads information about the items the user finds when browsing the recommender's web site. This information may include a description about the items, the ratings that other users assigned to them, and the categories of interest these items belong to. In Amazon, for instance, all this information is available to users.

However, as commented on in Sec. 4.1, this is not always the case. For this reason, our approach incorporates modules intended to retrieve the population's ratings and categorize all the items that the user explores.

On the other hand, this module receives the ratings sent by the *suppression alarm generator*. Afterwards, the module submits these ratings to the recommendation system.

Category Extractor. This component is responsible for obtaining the categories the items belong to. To this end, the module uses the information provided by the communication manager. Should this information not be enough, the module will have to get additional data by searching the Web or by querying an information provider. Afterwards, the categorization of these items is carried out by using the vector space model and the TF-IDF weights as commented on in Sec. 4.1. In a last stage, this module sends the items and their corresponding categories to the user. If the user has an opinion about these items, then the user proceeds to rate them.

User Profile Constructor. This module is responsible for the estimation of the user profile. To this end, the module is provided with the items that the user rates, i.e., those items capturing their preferences. Based on the received items, this block generates the user profile as described in Sec. 4.1. Note that the items received by this module are not necessarily those ultimately submitted to the recommender—after rating each item, the user is advised on the suitability of sending it to the system. Obviously, during this process, the module discards those rated items that were already considered in the histogram computation.

Population Profile Constructor. This module is responsible for the estimation of the population's profile. For this purpose, the block continuously receives items captured by the communication manager. Alternatively, this block could query databases containing this kind of information. This would be the case, for example, of a future application similar to *Google Insight*.

Suppression Strategy Generator. This block is the centrepiece of the architecture as it is directly responsible for the user privacy. First, the block is provided with the user profile and the population's profile. In addition, the user specifies a suppression rate σ, which is the relative frequency of ratings that the user is disposed to eliminate. Having specified this rate, the module computes the optimum tuple of suppression r^*, which contains information about the ratings that should be suppressed. More accurately, the component r_i is the percentage of ratings to items that our architecture suggests eliminating in the category i. An example of this is represented in Fig. 4, where we suppose that the user agrees to eliminate $\sigma = 15\%$ of their ratings. Based on this rate, the block calculates the optimal tuple r^*. In this example, the tuple r^* indicates that the user should refrain from rating 5% of the items belonging to the category 1 and 10% in the category 2. This is consistent with the fact that the actual user profile slightly deviates from the population's profile in these categories.

In the end, this tuple is sent to the suppression alarm generator. Later in Sec. 6, we provide a more detailed specification of this module by using a formulation of the trade-off between privacy and suppression rate, which will enable us to compute the tuple r^*.

Suppression Alarm Generator. This module is responsible for warning the user when their privacy is being compromised. Concretely, this module receives the tuple r^*. When the user decides to assign a rating to one of the items categorized by the category extractor, the module proceeds as follows. First, this item is sent to the user profile constructor, which updates the profile. Secondly, if r^* has a positive component in at least one of the categories

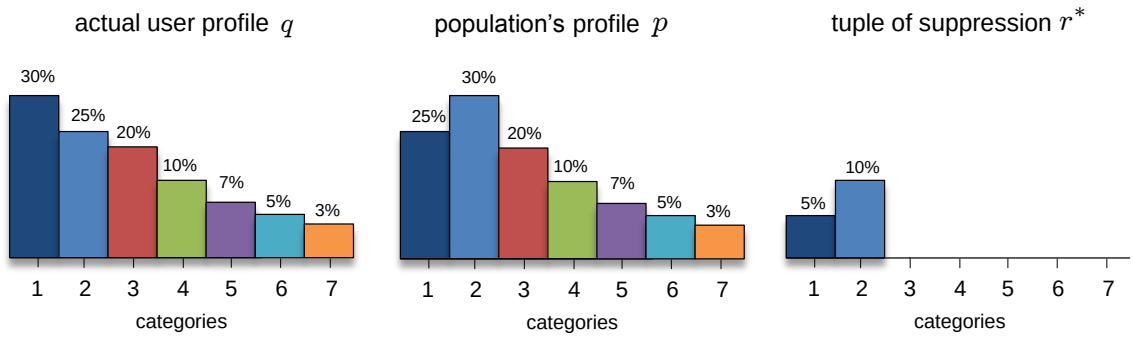

Figure 4. Here we illustrate an example in which a user with profile q is disposed to eliminate $\sigma = 15\%$ of their total number of ratings. Based on the user profile, the average population profile and the suppression rate, our approach computes the optimal tuple r^*, which provides the user with the proportion of items that they should eliminate in each category.

the item belongs to, a privacy alarm is generated to alert the user, and it is then for the user to decide whether to ultimately eliminate the rating or not. However, if r^* is zero for all components, our architecture does not become aware of any privacy risk and the rating is sent to the communication manager module. This process is repeated provided that the user attempts to rate an item they have an opinion on.

In order to illustrate how this block works, suppose that it receives the tuple of suppression shown in Fig. 4. According to this, the block would trigger an alarm if the user decided to rate an item classified into the categories 1 or 2. On the contrary, if the user wanted to rate an item belonging to any of the other categories, the system would forward this rating to the recommender.

After having explored each of the modules of the architecture, next we shall describe how our approach would work. Initially, the user would browse the recommendation system's web site and would find some items. In order for the user to obtain future recommendations from the system, they would have to rate some of those items. Before proceeding, though, our approach would retrieve information about the items and extract the categories they belong to. Afterwards, the user would try to rate one of those items, what would update the user profile and allow our system to compute the tuple r^*. In the end, our approach could suggest refraining from rating the item. Should this be the case, the user would have to decide whether to send the rating or not.

5 Justification of Entropy and Divergence as Measures of Privacy

In this section, we shall justify and interpret the privacy metric considered in our approach, already introduced in Sec. 4.4. Since KL divergence is a generalization of Shannon's entropy, we shall find that our interpretation may also be extended to entropy as a measure of privacy. For that purpose, we shall adopt the perspective of Jaynes' celebrated *rationale on entropy maximization methods* [44], which builds upon the *method of types* [36, §11], a powerful technique in large deviation theory whose fundamental results we proceed to review.

The first part of this section will tackle an important question. Suppose we are faced with a problem, formulated in terms of a model, in which a probability distribution plays a major role. In the event this distribution is unknown, we wish to assume a feasible candidate. What is the most likely probability distribution? In other words, what is the "probability of a probability" distribution? We shall see that a widespread answer to this question relies on choosing the distribution *maximizing the Shannon entropy*, or, if a reference distribution is available, the distribution *minimizing the KL divergence* with respect to it, commonly subject to feasibility constraints determined by the specific application at hand.

Our review of the maximum entropy method is crucial because it is unfortunately not always known in the privacy community, and because the rest of this paper constitutes a sophisticated illustration of its application, in the context of the protection of the privacy of user profiles. As we shall see in the second part of this section, the key idea is to model a user profile as a histogram of relative frequencies across categories of interest, regard it as a probability distribution, apply the maximum entropy method to measure the likelihood of a user profile either as its entropy or as its divergence with respect to the population's average profile, and finally take that likelihood as a measure of anonymity.

5.1 Rationale behind the Maximum Entropy Method

A wide variety of models across diverse fields have been explained on the basis of the intriguing principle of entropy maximization. A classical example in physics is the Maxwell-Boltzmann probability distribution $p(v)$ of particle velocities V in a gas [45, 46] of known temperature. It turns out that $p(v)$ is precisely the probability distribution maximizing the entropy, subject to a constraint on the temperature, equivalent to a constraint on the average kinetic energy, in turn equivalent to a constraint on $\mathrm{E}\,V^2$. Another well-known example, in the field of electrical engineering, of the application of the maximum entropy method, is Burg's spectral estimation method [47]. In this method, the power spectral density of a signal is regarded as a probability distribution of power across frequency, only partly known. Burg suggested filling in the unknown portion of the power spectral density by choosing that maximizing the entropy, constrained on the partial knowledge available. More concretely, in discrete case, when the constraints consist in a given range of the crosscorrelation function, up to a time shift k, the solution turns out to be a k^{th} order Gauss-Markov process [36]. A third and more recent example, this time in the field of natural language processing, is the use of log-linear models, which arise as the solution to constrained maximum entropy problems [48] in computational linguistics.

Having motivated the maximum entropy method, we are ready to proceed to describe Jaynes' attempt to justify, or at least interpret it, by reviewing the method of types of large deviation theory, a beautiful area lying at the intersection of statistics and information theory. Let $X_1, \ldots, X_k$ be a sequence of k i.i.d. drawings of an r.v. uniformly distributed in the alphabet $\{1, \ldots, n\}$. Let k_i be the number of times symbol $i = 1, \ldots, n$ appears in a sequence of outcomes $x_1, \ldots, x_k$, thus $k = \sum_i k_i$. The *type* t of a sequence of outcomes is the relative proportion of occurrences of each symbol, that is, the *empirical distribution* $t = \left(\frac{k_1}{k}, \ldots, \frac{k_n}{k}\right)$, not necessarily uniform. In other words, consider tossing an n-sided fair dice k times, and seeing exactly k_i times face i. In [44], Jaynes points out that

$$\mathrm{H}(t) = \mathrm{H}\left(\frac{k_1}{k}, \ldots, \frac{k_n}{k}\right) \simeq \frac{1}{k} \log \frac{k!}{k_1! \cdots k_n!} \quad \text{for } k \gg 1.$$

Loosely speaking, for large k, the size of a *type class*, that is, the number of possible outcomes for a given type t (permutations with repeated elements), is approximately $2^{k\,\mathrm{H}(t)}$ in the exponent. The fundamental rationale in [44] for selecting the type t with maximum entropy $\mathrm{H}(t)$ lies in the approximate equivalence between entropy maximization and the maximization of the number of possible outcomes corresponding to a type. In a way, this justifies the infamous *principle of insufficient reason*, according to which, one may expect an approximately equal relative frequency $k_i/k = 1/n$ for each symbol i, as the uniform distribution maximizes the entropy. The principle of entropy maximization is extended to include constraints also in [44].

Obviously, since all possible permutations count equally, the argument only works for uniformly distributed drawings, which is somewhat circular. A more general argument [36, §11], albeit entirely analogous, departs from a prior knowledge of an arbitrary PMF $\bar{t}$, not necessarily uniform, of such samples $X_1, \ldots, X_k$. Because the empirical distribution or type T of an i.i.d. drawing is itself an r.v., we may define its PMF $p(t) = \mathrm{P}\{T = t\}$; formally, the PMF of a random PMF. Using indicator r.v.'s, it is straightforward to confirm the intuition that $\mathrm{E}\,T = \bar{t}$. The general argument in question leads to approximating the probability $p(t)$ of a type class, a fractional measure of its size, in terms of its relative entropy, specifically $2^{-k\,\mathrm{D}(t\,\|\,\bar{t})}$ in the exponent, i.e.,

$$\mathrm{D}(t\,\|\,\bar{t}) \simeq -\frac{1}{k}\,\log p(t) \quad \text{for } k \gg 1,$$

which encompasses the special case of entropy, by virtue of (1). Roughly speaking, the likelihood of the empirical distribution t exponentially decreases with its KL divergence with respect to the average, reference distribution $\bar{t}$.

In conclusion, the most likely PMF t is that minimizing its divergence with respect to the reference distribution $\bar{t}$. In the special case of uniform $\bar{t} = u$, this is equivalent to maximizing the entropy, possibly subject to constraints on t that reflect its partial knowledge or a restricted set of feasible choices. The application of this idea to justify the privacy criterion assumed in our approach is the object of the remainder of this section.

5.2 Measuring the Privacy of User Profiles

We are finally equipped to justify, or at least interpret, our proposal to adopt Shannon's entropy and KL divergence as measures of the privacy of a user profile. Before we dive in, we must stress that the use of entropy as a measure of privacy, in the widest sense of the term, is by no means new. Shannon's work in the fifties introduced the concept of *equivocation* as the conditional entropy of a private message given an observed cryptogram [49], later used in the formulation of the problem of the wiretap channel [50, 51] as a measure of confidentiality. More recent studies [33, 34] rescue the suitable 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. More recent work has taken initial steps in relating privacy to information-theoretic quantities [11, 30–32].

In the context of this paper, an intuitive justification in favor of entropy maximization is that it boils down to making the apparent user profile as uniform as possible, thereby hiding a user's particular bias towards certain categories of interest. But a much richer argumentation stems from Jaynes' rationale behind entropy maximization methods [44, 52],

more generally understood under the beautiful perspective of the method of types and large deviation theory [36, §11], which we motivated and reviewed in the previous subsection.

Under Jaynes' rationale on entropy maximization methods, the entropy of an apparent user profile, modeled by a relative frequency histogram of categorized queries, may be regarded as a measure of privacy, or perhaps more accurately, anonymity. The leading idea is that the method of types from information theory establishes an approximate monotonic relationship between the likelihood of a 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 PMFs, viewed abstractly as r.v.'s 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 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, less interesting to an attacker assumed to strive to target peculiar users.

If an aggregated histogram of the population were available as a reference profile, the extension of Jaynes' argument to relative entropy, that is, to the KL divergence, would also give an acceptable measure of privacy (or anonymity). Note that is precisely the assumption made in the architecture described in Sec. 4, where the population's profile is available to users. Recall from Sec. 3 that KL divergence is a measure of discrepancy between probability distributions, which includes Shannon's entropy as the special case when the reference distribution is uniform. Conceptually, a lower KL divergence hides discrepancies with respect to a reference profile, say the population's, and there also exists a monotonic relationship between the likelihood of a distribution and its divergence with respect to the reference distribution of choice, which enables us to regard KL divergence as a measure of anonymity in a sense entirely analogous to the above mentioned. In fact, KL divergence was used recently in our own work [11,35] as a generalization of entropy to measure privacy, although the justification used built upon a number of technicalities, and the connection to Jaynes' rationale was not nearly as detailed as in this manuscript.

6 Formulation of the Trade-Off Privacy and Suppression Rate

In this section, we present a formulation of the optimal trade-off between privacy and suppression rate. In the absence of a thorough study, our formulation considers this rate as a measure of the degradation in the accuracy of the recommendations. This simplification allows us to formulate the problem of choosing a suppression tuple as a multiobjective optimization problem that takes into account privacy and suppression rate. As we shall show later, this formulation will enable us to go into the details of one of the functional blocks of the architecture described in Sec. 4.4.

Next, we formalize some of the concepts that we introduced in Sec. 4. Specifically, we model the *items* in a recommendation system as r.v.'s taking on values in a common finite alphabet of categories, namely the set $\{1, \ldots, n\}$ for some $n \in \mathbb{Z}^+$. Accordingly, we define q as the probability distribution of the items a *user* has an opinion on, that is, the distribution capturing the actual preferences of the user. In line with Sec. 4.4, we introduce a *rating suppression* rate $\sigma \in [0, 1)$, modeling the proportion of items that the user consents to eliminate. Bearing this in mind, we define the user's *apparent* item distribution s as $\frac{q-r}{1-\sigma}$

for some suppression strategy $r = (r_1, \ldots, r_n)$ satisfying $q_i \geqslant r_i \geqslant 0$ and $\sum r_i = \sigma$ for $i = 1, \ldots, n$. In light of this definition, the user's apparent item distribution may be interpreted as the result of the suppression of some ratings of items and the posterior normalization by $\frac{1}{1-\sigma}$ so that $\sum_i s_i = 1$.

Taking into account the definition of our privacy criterion, justified previously in Sec. 5, we shall suppose that the population is large enough to neglect the impact of the choice of r on p. Accordingly, we define the *privacy-suppression* function

$$\mathcal{R}(\sigma) = \min_{\substack{r \\ q_i \geqslant r_i \geqslant 0, \sum r_i = \sigma}} \mathrm{D}\left(\frac{q-r}{1-\sigma} \,\|\, p\right), \tag{2}$$

which poses the optimal trade-off between privacy (risk) and suppression rate and enables us to specify the module *suppression strategy generator* in Sec. 4.4. More accurately, this functional block will be in charge of solving the optimization problem (2).

There are two important advantages in modeling the privacy of a user profile as a divergence in general, or an entropy in particular, in this and other potential applications of our privacy criterion. First, the mathematical tractability demonstrated in [11]. Secondly, the privacy-suppression function has been defined in terms of an optimization problem, whose objective function is convex, subject to an affine constraint. As a consequence, this problem belongs to the extensively studied class of convex optimization problems [53] and may be solved numerically, using a number of extremely efficient methods, such as interior-point methods.

7 Concluding Remarks

There exist numerous proposals for the protection of user privacy in recommendation systems. Within those approaches, the suppression of ratings arises as a simple mechanism in terms of infrastructure requirements, as users need not trust the recommender. Nonetheless, the application of this privacy-enhancing technique comes at the cost of some processing overhead and, more importantly, at the expense of a degradation in the accuracy of the recommendations.

Our first contribution is an architecture that implements the suppression of ratings in those recommendation systems that profile users exclusively from their ratings. We describe the functionality of the internal modules of this architecture. The centrepiece of our approach is a module responsible for computing a tuple containing information about which ratings should be eliminated. Our architecture uses then this information to warn the user when their profile diverges from the population's rating distribution. The user is who finally decides whether to follow the recommendations made by our approach or not.

Privacy risk is measured as the KL divergence between the user's rating distribution and the population's, a criterion that we proposed in previous work [11] for query forgery in PIR. The justification of this criterion in the scenario of recommendation systems is our second contribution. First, we thoroughly interpret this metric by elaborating on the intimate connection between the celebrated method of entropy maximization and the use of entropies and divergences as measures of privacy. Measuring privacy enables us to optimize it, drawing upon powerful tools of convex optimization. The entropy maximization method is a beautiful principle widely used in fields such as physics, electrical engineering and even natural language processing. Secondly, we attempt to bridge the gap between the privacy

and the information-theoretic communities by substantially adapting some technicalities of our original work to reach a wider audience, not intimately familiar with information theory and the method of types. As neither information theory nor convex optimization are fully widespread in the privacy community, we elaborate and clarify the connection with privacy in far more detail, and hopefully in more accessible terms, than in our original work.

Lastly, we present a mathematical formulation of the optimal trade-off between privacy and suppression rate, which arises from the definition of our privacy criterion. This formulation allows us to specify the module responsible for user privacy in our architecture.

Acknowledgments

This work was partly supported by the Spanish Government through projects Consolider Ingenio 2010 CSD2007-00004 "ARES", TEC2010-20572-C02-02 "Consequence" and by the Government of Catalonia under grant 2009 SGR 1362. David Rebollo-Monedero is the recipient of a Juan de la Cierva postdoctoral fellowship, JCI-2009-05259, from the Spanish Ministry of Science and Innovation.

References

[1] "Amazon.com." [Online]. Available: http://www.amazon.com

[2] "Movielens." [Online]. Available: http://movielens.umn.edu

[3] "Netflix." [Online]. Available: http://www.netflix.com

[4] "Digg." [Online]. Available: http://digg.com

[5] D. Oard and J. Kim, "Implicit feedback for recommender systems," in *Proc. AAAI Workshop Recommender Syst.*, 1998, pp. 81–83.

[6] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, Dec. 1992.

[7] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, Jan. 2009.

[8] L. F. Cranor, ""I didn't buy it for myself". Privacy and e-commerce personalization," in *Proc. Workshop Priv. Electron. Society*, Washington, DC, 2003, pp. 111–117.

[9] J. Zaslow, "If TiVo thinks you are gay, here's how to set it straight," Nov. 2002. [Online]. Available: http://online.wsj.com/article_email/SB1038261936872356908.html

[10] D. L. Hoffman, T. P. Novak, and M. Peralta, "Building consumer trust online," *Commun. ACM*, vol. 42, no. 4, pp. 80–85, Apr. 1999.

[11] D. Rebollo-Monedero and J. Forné, "Optimal query forgery for private information retrieval," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4631–4642, 2010.

[12] H. Polat and W. Du, "Privacy-preserving collaborative filtering using randomized perturbation techniques," in *Proc. SIAM Int. Conf. Data Min. (SDM)*. IEEE Comput. Soc., 2003.

[13] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*. Washington, DC: IEEE Comput. Soc., 2003, pp. 99–106.

[14] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*. ACM, 2005, pp. 37–48.

[15] H. Polat and W. Du, "SVD-based collaborative filtering with privacy," in *Proc. ACM Int. Symp. Appl. Comput. (SASC)*. ACM, 2005, pp. 791–795.

[16] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Santa Barbara, CA, 2001, pp. 247–255.

[17] "Jester: The online joke recommender." [Online]. Available: http://eigentaste.berkeley.edu/

[18] J. Canny, "Collaborative filtering with privacy via factor analysis," in *Proc. ACM SIGIR Conf. Res., Develop. Inform. Retrieval*. Tampere, Finland: ACM, 2002, pp. 238–245.

[19] J. F. Canny, "Collaborative filtering with privacy," in *Proc. IEEE Symp. Secur., Priv. (SP)*, 2002, pp. 45–57.

[20] W. Ahmad and A. Khokhar, "An architecture for privacy preserving collaborative filtering on web portals," in *Proc. IEEE Int. Symp. Inform. Assurance, Secur. (IAS)*. Washington, DC: IEEE Comput. Soc., 2007, pp. 273–278.

[21] J. Zhan, C. L. Hsieh, I. C. Wang, T. S. Hsu, C. J. Liau, and D. W. Wang, "Privacy-preserving collaborative recommender systems," *IEEE Trans. Syst. Man, Cybern.*, vol. 40, no. 4, pp. 472–476, Jul. 2010.

[22] B. Miller, N. Bradley, and J. A. K. J. Riedl, "Pocketlens: Toward a personal recommender system," *ACM Trans. Inform. Syst.*, vol. 22, no. 3, pp. 437–476, Jul. 2004.

[23] S. Berkovsky, Y. Eytani, T. Kuflik, and F. Ricci, "Enhancing privacy and preserving accuracy of a distributed collaborative filtering," in *Proc. ACM Conf. Recommender Syst. (RecSys)*. ACM, 2007, pp. 9–16.

[24] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression," SRI Int., Tech. Rep., 1998.

[25] X. Sun, H. Wang, J. Li, and T. M. Truta, "Enhanced p-sensitive k-anonymity models for privacy preserving data publishing," *Trans. Data Priv.*, vol. 1, no. 2, pp. 53–66, 2008.

[26] T. M. Truta and B. Vinay, "Privacy protection: p-sensitive k-anonymity property," in *Proc. Int. Workshop Priv. Data Manage. (PDM)*, Atlanta, GA, 2006, p. 94.

[27] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian, "l-Diversity: Privacy beyond k-anonymity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Atlanta, GA, Apr. 2006, p. 24.

[28] H. Jian-min, C. Ting-ting, and Y. Hui-qun, "An improved V-MDAV algorithm for l-diversity," in *Proc. IEEE Int. Symp. Inform. Process. (ISIP)*, Moscow, Russia, May 2008, pp. 733–739.

[29] J. Domingo-Ferrer and V. Torra, "A critique of k-anonymity and some of its enhance-

ments," in *Proc. Workshop Priv., Secur., Artif. Intell. (PSAI)*, Barcelona, Spain, 2008, pp. 990–993.

[30] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Istanbul, Turkey, Apr. 2007, pp. 106–115.

[31] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From t-closeness to PRAM and noise addition via information theory," in *Priv. Stat. Databases (PSD)*, ser. Lecture Notes Comput. Sci. (LNCS). Istambul, Turkey: Springer-Verlag, Sep. 2008, pp. 100–112.

[32] ——, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190

[33] C. Díaz, S. Seys, J. Claessens, and B. Preneel, "Towards measuring anonymity," in *Proc. Workshop Priv. Enhanc. Technol. (PET)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 2482. Springer-Verlag, Apr. 2002, pp. 54–68.

[34] C. Díaz, "Anonymity and privacy in electronic services," Ph.D. dissertation, Katholieke Univ. Leuven, Dec. 2005.

[35] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "A privacy-preserving architecture for the semantic web based on tag suppression," in *Proc. Int. Conf. Trust, Priv., Secur., Digit. Bus. (TRUSTBUS)*, Bilbao, Spain, Aug. 2010, pp. 58–68.

[36] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[37] G. Bianchi, M. Bonola, V. Falletta, F. S. Proto, and S. Teofili, "The SPARTA pseudonym and authorization system," *Sci. Comput. Program.*, vol. 74, no. 1–2, pp. 23–33, 2008.

[38] V. Benjumea, J. López, and J. M. T. Linero, "Specification of a framework for the anonymous use of privileges," *Telemat., Informat.*, vol. 23, no. 3, pp. 179–195, Aug. 2006.

[39] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy preserving targeted advertising," in *Proc. IEEE Symp. Netw. Distrib. Syst. Secur. (SNDSS)*, 2010, pp. 1–21.

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

[41] J. Domingo-Ferrer, "Coprivacy: Towards a theory of sustainable privacy," in *Priv. Stat. Databases (PSD)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 6344. Corfu, Greece: Springer-Verlag, Sep. 2010, pp. 258–268.

[42] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[43] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[44] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, Sep. 1982.

[45] L. Brillouin, *Science and Information Theory.* New York: Academic-Press, 1962.

[46] E. T. Jaynes, *Papers on Probability, Statistics and Statistical Physics.* Dordrecht: Reidel, 1982.

[47] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. dissertation, Stanford Univ., 1975.

[48] A. L. Berger, J. della Pietra, and A. della Pietra, "A maximum entropy approach to natural language processing," *MIT Comput. Ling.*, vol. 22, no. 1, pp. 39–71, Mar. 1996.

[49] C. E. Shannon, "Communication theory of secrecy systems," Bell Syst., Tech. J., 1949.

[50] A. Wyner, "The wiretap channel," Bell Syst., Tech. J. 54, 1975.

[51] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. Inform. Theory*, vol. 24, pp. 339–348, May 1978.

[52] E. T. Jaynes, "Information theory and statistical mechanics II," *Phys. Review Ser. II*, vol. 108, no. 2, pp. 171–190, 1957.

[53] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge, UK: Cambridge University Press, 2004.