Privacy Tradeoffs in Predictive Analytics

Stratis Ioannidis*, Andrea Montanari†, Udi Weinsberg*,
Smriti Bhagat*, Nadia Fawaz*, Nina Taft*
* Technicolor, † Stanford University

March 31, 2014

Abstract

Online services routinely mine user data to predict user preferences, make recommendations, and place targeted ads. Recent research has demonstrated that several private user attributes (such as political affiliation, sexual orientation, and gender) can be inferred from such data. Can a privacy-conscious user benefit from personalization while simultaneously protecting her private attributes? We study this question in the context of a rating prediction service based on matrix factorization. We construct a protocol of interactions between the service and users that has remarkable optimality properties: it is *privacy-preserving*, in that no inference algorithm can succeed in inferring a user's private attribute with a probability better than random guessing; it has *maximal accuracy*, in that no other privacy-preserving protocol improves rating prediction; and, finally, it involves a *minimal disclosure*, as the prediction accuracy strictly decreases when the service reveals less information. We extensively evaluate our protocol using several rating datasets, demonstrating that it successfully blocks the inference of gender, age and political affiliation, while incurring less than 5% decrease in the accuracy of rating prediction.

1 Introduction

Online users are routinely asked to provide feedback about their preferences and tastes. Often, users give five-star ratings for movies, books, restaurants, or items they purchase, and "like" news articles, blog posts, pictures or other kinds of micro-content. Online services mine such feedback to predict users' future preferences, using techniques such as matrix factorization [9,22–24]. Such prediction can be employed to, e.g., make relevant product recommendations, to display targeted ads, or, more generally, personalize services offered; making accurate predictions is thus of fundamental importance to many online services.

Although users may willingly reveal, e.g., ratings to movies or "likes" to news articles and posts, there is an inherent privacy threat in this revelation. To see this, consider the following general setting. An entity, which we call for concreteness the *analyst*, has a dataset of ratings given by users to a set of items (e.g., movies). A private attribute of some users, such as their gender, age, political affiliation, etc., is also in the dataset. The analyst uses this dataset to offer a *recommendation service.* Specifically, the analyst solicits ratings from new users; using these ratings, it predicts how these users would rate other items in its dataset (e.g., via matrix factorization techniques), and recommends items they are likely to rate highly. New users are *privacy-conscious*: they want to receive relevant recommendations but do not want the analyst to learn their private attribute. However, having access to the above dataset, the analyst can potentially *infer* the private attribute from the ratings they reveal.

The success of such inference clearly depends on how a user's feedback (i.e., her ratings) relates to her private attribute, and whether this correlation is evident in the dataset. Recent studies report many examples where strong correlations have been found: attributes successfully inferred from ratings or "likes" include political affiliation [25,38], sexual orientation [25], age [44], gender [38,44], and even drug use [25]. Yet more privacy threats have been extensively documented in literature (see, e.g., [3, 30, 31, 34, 35]). It is therefore natural to ask *how can a privacy-conscious user benefit from relevant recommendations, while preventing the inference of her private information*? Allowing this to happen is clearly desirable from the user's point of view. It also benefits the analyst, as it incentivizes privacy-conscious individuals to use the recommendation service.

A solution proposed by many recent research efforts is to allow a user to distort her ratings before revealing them to the analyst [7,12,21,42]. This leads to a well-known tradeoff between *accuracy* and *privacy*: greater distortion yields better privacy but also less accurate prediction (and, hence, poorer recommendations). We introduce for the first time a third dimension to this tradeoff, namely the *information the analyst discloses to the users.*

To understand the importance of this dimension, consider the following hypothetical scenario. The analyst gives the privacy-conscious user an implementation of its rating prediction algorithm, as well as any data it requires–including, potentially, the full dataset at the analyst's disposal. The user can then execute this algorithm locally, identifying, e.g., which movies or news articles are most relevant to her. This would provide perfect privacy (as the user reveals nothing to the analyst) as well as maximal accuracy (since the user's ratings are not distorted). Clearly, this is untenable from the analyst's perspective, both for practical reasons (e.g., efficiency or code maintenance) and for commercial reasons: the analyst may be charging a fee for its services, and exposing such information publicly diminishes any competitive edge it may have.

The above hypothetical scenario illustrates that *both* privacy *and* accuracy can be trivially attained when *no constraints are placed on the information disclosed by the analyst.* On the other hand, such constraints are natural and necessary when the analyst's algorithms and data are proprietary. A natural goal is thus to determine the *minimal* information the analyst needs to disclose to a privacy-conscious user, to enable a recommendation service that is both private and accurate. We make the following contributions:

- We introduce a novel mathematical framework to study issues of privacy, accuracy, and information disclosure when the analyst predicts ratings through matrix factorization (Section 4). In particular, we define a broad class of *learning protocols* determining the interactions between the analyst and a privacy-conscious user. Each protocol specifies what information the analyst reveals, how the user distorts her ratings, and how the analyst uses this obfuscated feedback for rating prediction.

- We propose a simple learning protocol, which we call the *midpoint* protocol (MP), and prove it has remarkable optimality properties (Section 5). First, it provides *perfect privacy* w.r.t. the user's private attribute: no inference algorithm predicts it better than random guessing. Second, it yields *optimal accuracy*: there is no privacy-preserving protocol allowing rating prediction at higher accuracy than MP. Finally, the protocol involves a *minimal disclosure*: any privacy-preserving protocol that discloses less information than MP necessarily has a strictly worse prediction accuracy.
- We extend our solution to handle common situations that occur in practice (Section 6). We deal with the case where the user can only rate a subset of the items for which the analyst solicits feedback: we provide a variant of MP, termed MPSS, and also establish its optimality in this setting. We also discuss how the analyst can select the set of items for which to solicit ratings, and how the user can repeatedly interact with the analyst.
- We evaluate our proposed protocols on three datasets, protecting attributes such as user gender, age and political affiliation (Section 7). We show that MP and MPSS attain excellent privacy: a wide array of inference methods are rendered no better than blind guessing, with an area-under-the-curve (AUC) below 0.55. This privacy is achieved with negligible impact (2-5%) on rating prediction accuracy.

To the best of our knowledge, we are the first to take into account the data disclosed by an analyst in the above privacy-accuracy tradeoff, and to establish the optimality of a combined disclosure, obfuscation, and prediction scheme. Our proofs rely on the modeling assumption that is the cornerstone of matrix factorization techniques and hence validated by vast empirical evidence (namely, that the user-item ratings matrix is approximately low-rank). Moreover, the fact that our algorithms successfully block inference against a barrage of different classifiers, some non-linear, further establishes our assumption's validity over real-world data.

2 Related Work

Threats. Inference threats from user data have been extensively documented by several recent studies. Demographic information has been successfully inferred from blog posts [3], search queries [5], reviews [34], tweets [35], and the profiles of one's Facebook friends [30]. In an extreme case of inference, Narayanan *et al.* [31] show that disclosure of movie ratings can lead to full deanonymization (through a linkage attack), thus enabling unique identification of users. Closer to our setting, Kosinski *et al.* [25] show that several personality traits, including political views, sexual orientation, and drug use can be accurately predicted from Facebook "likes", while Weinsberg *et al.* [44] show that gender can be inferred from movie ratings with close to 80% accuracy. Salamatian *et al.* [38] also show that political views can be inferred with confidence above 71% by using only a user's ratings to her 5 most-watched TV shows.

Privacy-Preserving Data Mining and Information-Theoretic Models. Distorting data prior to its release to an untrusted analyst has a long history in the context of privacy-preserving data mining (see, e.g., [42, 43]). Distortion vs. estimation accuracy tradeoffs have been studied in the context of several statistical tasks, such as constructing decision trees [1], clustering [2, 33], and parameter estimation [12]. The outcome of such tasks amounts to learning an aggregate property from the distorted data of a user population. In contrast, we focus on estimating accurately a user profile to be used in matrix factorization, while keeping private any attribute she deems sensitive.

Our setting is closely related to the following information-theoretic problem [7, 45]. Consider two dependent random variables X and Y, where X is to be released publicly while Y is to be kept secret. To prevent inference of Y from the release, one can apply a distortion $f(X)$ on X; the goal is then to find the minimal distortion so that the mutual information between $f(X)$ and Y is below a threshold. This problem was originally addressed in the asymptotic regime [39, 45], while a series of recent works study it in a non-asymptotic setting [7, 28, 36, 38]. Broadly speaking, our work can be cast in this framework by treating a user's ratings as X, her private feature as Y, and employing a correlation structure between them as specified by matrix factorization (namely, (7)). Our definition of privacy then corresponds to zero mutual information (i.e., "perfect" privacy), and our protocol involves a minimal rating distortion.

We depart from these studies of privacy vs. accuracy (both in information-theoretic as well as the privacy-preserving data mining settings), by investigating a third axis, namely, the information disclosed by the analyst. To the best of our knowledge, our work is the first to characterize the disclosure extent necessary to achieve an optimal privacy-accuracy trade-off, an aspect absent from the aforementioned works.

Trusted Analyst. A different threat model than the one we study here considers a trusted analyst that aggregates data from multiple users in the clear. The analyst performs a statistical operation over the data, distorts the output of this operation, and releases it publicly. The privacy protection gained by the distortion is therefore towards a third party that accesses the distorted output. The most common approach to quantifying privacy guarantees in this setting is through *ϵ-differential privacy* [13, 15]. The statistical operations studied under this setting are numerous, including social recommendations [27], covariance computation [29], statistical estimation [14, 40], classification [10, 37], and principal component analysis [11], to name a few. We differ in considering an untrusted analyst, and enabling a privacy-conscious user to interact with an analyst performing matrix factorization, rather than learning aggregate statistics.

3 Technical Background

In this section, we briefly review matrix factorization and the modeling assumptions that underlie it. We also highlight privacy challenges that arise from its application.

3.1 Matrix Factorization (MF)

Matrix factorization [8, 22, 24] addresses the following prediction problem. A data analyst has access to a dataset in which N users rate subsets of M possible items (e.g., movies, restaurants, news articles, etc.). For $[N] \equiv \{1, \dots, N\}$ the set of users, and $[M] \equiv \{1, \dots, M\}$ the set of items, we denote by $\mathcal{E} \subseteq [N] \times [M]$ the user-item pairs with a rating in the dataset. For $(i, j) \in \mathcal{E}$, let $r_{ij} \in \mathbb{R}$ be user i's rating to item j. Given the dataset $\{(i, j, r_{i,j})\}_{(i,j)\in\mathcal{E}}$, the analyst wishes to predict the ratings for user-item pairs $(i, j) \notin \mathcal{E}$.

Matrix factorization attempts to solve this problem assuming that the $N \times M$ matrix comprising all ratings is *approximately low-rank*. In particular, it is assumed that for some small dimension $d \in \mathbb{N}$ there exist vectors $\mathbf{x}_i, \mathbf{v}_j \in \mathbb{R}^d$, termed the *user* and *item profiles*, respectively, such that

$$r_{ij} = \langle \mathbf{x}_i, \mathbf{v}_i \rangle + \varepsilon_{ij}, \quad \text{for } i \in [N], j \in [M], \tag{1}$$

where the "noise" variables ε_{ij} are zero mean, i.i.d. random variables with finite variance, and $\langle \mathbf{a}, \mathbf{b} \rangle \equiv \sum_{k=1}^{d} a_k b_k$ is the usual scalar product in $\mathbb{R}^d$. Given the ratings $\{r_{ij},\ (i, j) \in \mathcal{E}\}$, the user and item profiles are typically computed through the following least-squares estimation (LSE) [24]:

$$\min_{\{\mathbf{x}_i\}_{i\in[N]}, \{\mathbf{v}_j\}_{j\in[M]}} \textstyle\sum_{(i,j)\in\mathcal{E}} (r_{ij} - \langle \mathbf{x}_i, \mathbf{v}_j \rangle)^2. \tag{2}$$

Minimizing this square error is a natural objective. Moreover, when the noise variables in (1) are Gaussian, (2) is equivalent to maximum likelihood estimation of user and item profiles. Note that, having solved (2), the analyst can predict the rating of user i for item j as:

$$\hat{r}_{ij} \equiv \langle \mathbf{x}_i, \mathbf{v}_j \rangle, \quad (i, j) \notin \mathcal{E}. \tag{3}$$

where $\mathbf{x}_i, \mathbf{v}_j$ are the estimated profiles obtained from (2).

Unfortunately, the minimization (2) is *not* a convex optimization problem. Nevertheless, there exist algorithms that provably recover the correct user and item profiles, under appropriate assumptions [8,9,22]. Moreover, simple gradient descent or alternating least-squares techniques are known to work very well in practice [24].

3.2 Incorporating Biases

Beyond user ratings, the analyst often has additional "contextual" information about users in the dataset. For example, if users are not privacy-conscious, they may reveal features such as their gender, age or other demographic information along with their ratings. Such information is typically included in MF through *biases* (see, e.g., [23, 24]).

Suppose, for concreteness, that each user i discloses a binary feature $x_{i0} \in \{-1, +1\}$, e.g., their gender or political affiliation. This information can be incorporated in MF by adapting the model (1) as follows:

$$r_{ij} = \langle \mathbf{x}_i, \mathbf{v}_j \rangle + x_{i0} v_{j0} + \varepsilon_{ij} = \langle x_i, v_j \rangle + \varepsilon_{ij} \tag{4}$$

for all $i \in [N]$, $j \in [M]$, where $v_{j0} \in \mathbb{R}$ is a type-dependent bias, and $x_i = (x_{i0}, \mathbf{x}) \in \mathbb{R}^{d+1}$, $i \in [N]$, $v_j = (v_{j0}, \mathbf{v}_j) \in \mathbb{R}^{d+1}$, $j \in [M]$, are *extended* user and item profiles, respectively. Under this modeling assumption, the analyst can estimate profiles and biases jointly by solving:

$$\min_{\{\mathbf{x}_i\}_{i \in [N]}, \{(v_{j0}, \mathbf{v}_j)\}_{j \in [M]}} \textstyle\sum_{(i,j) \in \mathcal{E}} (r_{ij} - \langle x_i, v_j \rangle)^2. \tag{5}$$

Note that this minimization can be seen as a special case of (2), in which extended profiles have dimension $d+1$, and the first coordinate of x_i is fixed to either -1 or $+1$ (depending on the user's binary feature x_{i0}). In other words, the feature x_{i0} can be treated as yet another feature of a user's profile, though it is *explicit* (i.e., a priori known) rather than *latent* (i.e., inferred through MF). Prediction can be performed again through $\hat{r}_{ij} = \langle x_i, v_j \rangle$, for $(i, j) \notin \mathcal{E}$.

Intuitively, the biases v_{j0} gauge the impact of the binary feature x_{i0} on each user's ratings. Indeed, consider sampling a random user from a population, and let $x = (x_0, \mathbf{x})$ be her profile, where $\mathbf{x}$ comprises the features that are independent of x_0. Then, it is easy to check from (4) that her rating r_j for item j will be such that:

$$\mathrm{E}\{r_j \mid x_0 = 1\} - \mathrm{E}\{r_j \mid x_0 = 0\} = 2v_{j0},$$

where the expectation is over the noise in (4), as well as the random sampling of the user. Put differently, given access to ratings by users that are not privacy-conscious and have disclosed, e.g., their gender x_0, v_{j0} corresponds to half the distance between the mean ratings for item j among genders.

Additional explicit binary features can be incorporated similarly, by adding one bias per feature in (5) (see, e.g., [24]). Categorical features can also be added through binarization; for simplicity, we focus on a single binary feature, discussing multiple and categorical features in Section 6.4.

3.3 Prediction for Privacy-Conscious Users

Consider a scenario in which the analyst has performed MF over a dataset of users that disclose a binary feature, extracting thusly the extended profiles $v_j = (v_{j0}, \mathbf{v}_j) \in \mathbb{R}^{d+1}$ for each item $j \in [M]$. Suppose now a privacy-conscious user joins the system and *does not explicitly reveal her private binary feature x_0 to the analyst.*

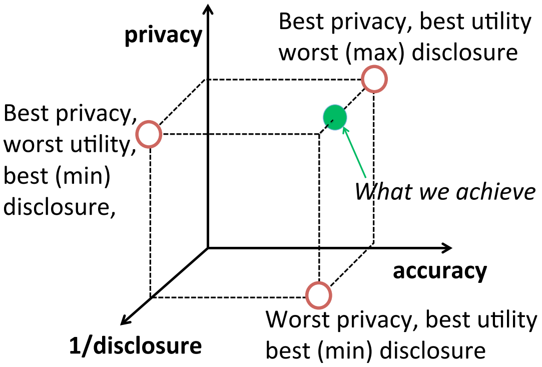


Figure 1: The red circles represent the three extreme protocols (Sec. 4.1) that fail to meet all three properties simultaneously. Our solution (see Sec. 5) lies near the upper front right edge of the cube, as it has perfect privacy and accuracy. We will prove (Thm. 1 in Sec. 5) that the region between our solution and the optimal corner ('zero' disclosure, perfect privacy, maximal accuracy) is unattainable.

In such a "cold-start" situation, the analyst would typically solicit a batch of ratings $\{r_j\}_{j\in S}$ for some set $\mathcal{S} \subseteq M$. Assume that the new user's ratings also follow the linear model (4) with extended profile $x = (x_0, \mathbf{x}) \in \{-1, +1\} \times \mathbb{R}^d$. Then, the analyst can (a) infer the user's extended profile x, and (b) predict her ratings for items in $[M] \setminus \mathcal{S}$ using the extended item profiles $\{v_j\}_{j\in\mathcal{S}}$ as follows. First, the analyst can infer x using through the LSE:

$$\min_{x_0\in\{-1,+1\},\mathbf{x}\in\mathbb{R}^d} \textstyle\sum_{j\in\mathcal{S}}(r_j - \langle \mathbf{x}, \mathbf{v}_j\rangle - x_0 v_{j0})^2. \tag{6}$$

The minimization (6) can be computed in time linear in $|\mathcal{S}|$, by solving two linear regressions (one for each $x_0 \in \{-1, +1\}$) and picking the solution $(x_0, \mathbf{x})$ that yields the smallest error (6). Having obtained an estimate of the extended profile x, the analyst can predict ratings as $\hat{r}_j = \langle x, v_j\rangle$, for $j \notin \mathcal{S}$.

Beyond this LSE approach, the analyst can use a different classification algorithm to first infer the private feature x_0, such as logistic regression or support vector machines (SVMs). We refer the reader to, *e.g.*, [44], for the description of several such algorithms and their application over real rating data. Having an estimate of x_0, the analyst can proceed to solve (6) w.r.t. $\mathbf{x}$ alone, which involves a single linear regression.

In both of the above approaches (joint LSE, or separate inference of x_0) the analyst infers the private feature x_0. Indeed, the LSE method (6) is known to predict information such as gender or age with an accuracy between 65–83% over real datasets [4]; separate inference of the private information (through, e.g., logistic regression or SVMs) leads to 84-86% accuracy [44]. As such, by revealing her ratings, the user also reveals x_0, albeit indirectly and unintentionally.

4 Modeling Privacy Tradeoffs

Section 3.3 illustrates that serving a privacy-conscious user is not straightforward: there is a tension between the user's privacy and the utility she receives. Accurate profiling allows correct rating prediction and enables relevant recommendations, at the cost of the inadvertent revelation of the user's private feature. It is thus natural to ask whether the user can benefit from accurate prediction *while preventing the inference of this feature*. We will provide both rigorous and empirical evidence that – perhaps surprisingly – this is possible. Specific features can be obfuscated without harming personalization. One of our main contributions is to identify that, beyond this privacy-utility tradeoff, there is in fact a third aspect to this problem: namely, how much information the analyst discloses to the user. In what follows, we present a framework that addresses these issues.

4.1 Problem Formulation

Motivated by Section 3.3, we consider a setting comprising the two entities we have encountered so far, an *analyst* and a *privacy-conscious user*. The analyst has access to a dataset of ratings collected from users that are not privacy-conscious, and have additionally revealed to the analyst a binary feature. By performing matrix factorization over this dataset, the analyst has extracted extended item profiles $v_j = (v_{j0}, \mathbf{v}_j) \in \mathbb{R}^{d+1}$, $j \in [M]$, for a set of M items.

The analyst solicits the ratings of the privacy-conscious user for a subset of items $\mathcal{S} \in [M]$. We again assume that the user is parametrized by an extended profile $x = (x_0, \mathbf{x}) \in \{-1, +1\} \times \mathbb{R}^d$, and that her ratings follow (4). The analyst's goal is to profile the user and identify items that the user might rate highly in $[M] \setminus \mathcal{S}$. The user is willing to aid the analyst in correctly profiling her; however, she is privacy-conscious w.r.t. her private feature x_0, and wishes to *prevent its inference*. We thus wish to design a *protocol* for exchanging information between the analyst and the user that has three salient properties; we state these here informally, postponing precise definitions until Section 4.3:

(a) At the conclusion of the protocol, the analyst estimates $\mathbf{x}$, the non-private component of x, *as accurately as possible*.

(b) The analyst *learns nothing* about the private feature x_0.

(c) The user learns *as little as possible* about the extended profile v_j of each item j.

To highlight the interplay between these three properties, we discuss here three "non-solutions", i.e., three protocols that fail to satisfy all three properties. First, observe that the "empty" protocol (no information exchange) clearly satisfies (b) and (c), but not (a): the analyst does not learn $\mathbf{x}$. Second, the protocol in which the user discloses her ratings to the analyst "in the clear", as in Section 3.3, satisfies (a) and (c) but not (b): it allows the analyst to estimate *both* $\mathbf{x}$ *and* x_0 through, e.g., the LSE (6).

Finally, consider the following protocol. The analyst discloses all item profiles v_j, $j \in \mathcal{S}$, to the user. Subsequently, the user estimates $\mathbf{x}$ locally, by solving the linear regression (6) over her ratings in $\mathcal{S}$, with her private feature x_0 fixed. The user concludes the protocol by sending the obtained estimate of $\mathbf{x}$ to the analyst. Observe that this protocol satisfies (a) and (b), but not (c). In particular, the user learns the extended profiles of all items in their entirety.

These protocols illustrate that it is simple to satisfy any two of the above properties, but not all three. Each of the three "non-solutions" above are in fact extrema among protocols constrained by (a)-(c): each satisfies two properties in the best way possible, while completely failing on the third. In the conceptual schematic of Figure 1 we illustrate where these three extreme protocols lie.

There is a clear motivation, from a practical perspective, to seek protocols satisfying all three properties. Property (a) ensures that, at the conclusion of the protocol, the analyst learns the non-private component of the user's profile, and can use it to suggest new items–benefiting thusly the user, and motivating the existence of this service. Property (b) ensures that a privacy-conscious user receives this benefit *without revealing her private feature*, thereby incentivizing her participation. Finally, property (c) limits the extent at which the item profiles $\{v_j\}_{j\in\mathcal{S}}$ are made publicly available. Indeed, the item profiles and the dataset from which they were constructed are proprietary information: disclosing them to any privacy-conscious user, as described by the last non-solution, would allow any user to offer the same service. More generally, it is to the analyst's interest to enable properties (a) and (b), thereby attracting privacy-conscious users, while limiting the disclosure of any proprietary information and its exposure to competition.

It is natural to ask what is the precise statement of "as accurately as possible", "learns nothing", and "as little as possible" in the above description of (a)-(c). We provide such formal definitions below.

4.2 A Learning Protocol

To formalize the notions introduced in properties (a)-(c) of Section 4.1, we describe in this section the interactions between the privacy-conscious user and the analyst in a more precise fashion. Recall that the user is parametrized by an extended profile $x = (x_0, \mathbf{x}) \in \{-1, +1\} \times \mathbb{R}^d$, and that her ratings follow (4); namely,

$$r_j = \langle \mathbf{x}, \mathbf{v}_j \rangle + x_0 v_{j0} + \varepsilon_j = \langle x, v_j \rangle + \varepsilon_j, \quad j \in [M] \tag{7}$$

where $v_j \in \mathbb{R}^{d+1}$, is the extended profile of item j, extracted through MF, and ε_j are i.i.d. zero mean random variables of variance $\sigma^2 < \infty$. We note that, unless explicitly stated, we *do not* assume that the noise variables ε_j are Gaussian; our results will hold with greater generality.

We assume that the set of items $\mathcal{S} \subseteq [M]$, for which ratings are solicited, is an arbitrary set chosen by the analyst[1]. We restrict our attention to items with

[1] We discuss how the analyst can select the items in $\mathcal{S}$ in Section 6.2.

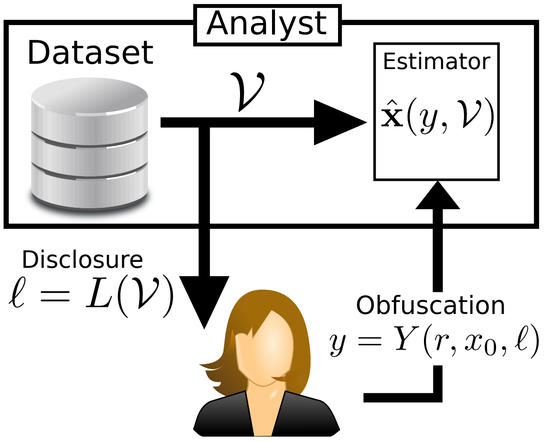


Figure 2: A learning protocol $\mathcal{R} = (L, Y, \widehat{\mathbf{x}})$ between an analyst and a privacy-conscious user. The analyst has access to a dataset, from which it extracts the extended profiles $\mathcal{V}$ through MF. It discloses to the user the information $\ell = L(\mathcal{V})$. Using this information, her vector of ratings r, and her private feature x_0, the user computes the obfuscated output $y = Y(r, x_0, \ell)$ and reveals it to the analyst. The latter uses this obfuscated feedback as well as the profiles $\mathcal{V}$ to estimate $\mathbf{x}$, using the estimator $\widehat{\mathbf{x}}(y, \mathcal{V})$.

extended profiles v_j such that $\mathbf{v}_j \neq \mathbf{0}$. Indeed, given the analyst's purpose of estimating $\mathbf{x}$, the rating of an item for which $\mathbf{v} = \mathbf{0}$ is clearly uninformative in light of (7). We denote by

$$\mathbb{R}^{d+1}_{-\mathbf{0}} \equiv \{(v_0, \mathbf{v}) \in \mathbb{R}^{d+1} : \mathbf{v} \neq \mathbf{0}\}$$

the set of all such vectors, and by $\mathcal{V} \equiv \{v_j,\ j \in \mathcal{S}\} \subseteq \mathbb{R}^{d+1}_{-\mathbf{0}}$ the extended profiles of items in $\mathcal{S}$. Recall that the user does not a priori know $\mathcal{V}$. In addition, the user knows her private variable x_0 and either knows or can easily generate her rating r_j to each item $j \in \mathcal{S}$. Nevertheless, the user *does not a-priori know* the remaining profile $\mathbf{x} \in \mathbb{R}^d$. This is consistent with MF, as the "features" corresponding to each coordinate of $\mathbf{v}_j$ are "latent".

The assumption that the user either knows or can easily generate her ratings in $\mathcal{S}$ is natural when the user can immediately form an opinion (this is the case for items such as blog posts, ads, news articles, tweets, pictures, short videos, etc.); or, when the "rating" is automatically generated from user engagement (e.g., it is the time a user spends at a website, or the reading collected by a skin sensor); or, when the user is obligated to generate a response (because, e.g., she is paid to do so). We discuss the case where the user can readily produce ratings for only a subset of $\mathcal{S}$ in Section 6.1.

Using the above notation, we define a *privacy-preserving learning protocol* as a protocol consisting of the following three components, as illustrated in Figure 2:

Disclosure. The disclosure determines the amount of information that the analyst discloses to the user regarding the profiles in $\mathcal{V}$. Formally, it is a mapping

$$L : \mathbb{R}^{d+1}_{-\mathbf{0}} \to \mathcal{L},$$

with $\mathcal{L}$ a generic set[2]. This is implemented as a program and executed by the analyst, who discloses to the user the information $\ell_j \equiv L(v_j) \in \mathcal{L}$ for each item $j \in \mathcal{S}$. We denote by $L(\mathcal{V})$ the vector $\ell \in \mathcal{L}^{|\mathcal{S}|}$ with coordinates ℓ_j, $i \in \mathcal{S}$. We note that, in practice, $L(\mathcal{V})$ can be made public, as it is needed by all potential privacy-conscious users that wish to interact with the analyst.

Obfuscation Scheme. The obfuscation scheme describes how user ratings are modified (obfuscated) before being revealed to the analyst. Formally, this is a mapping

$$Y : \mathbb{R}^{|\mathcal{S}|} \times \{-1, +1\} \times \mathcal{L}^{|\mathcal{S}|} \to \mathcal{Y},$$

for $\mathcal{Y}$ again a generic set. The mapping is implemented as a program and executed by the user. In particular, the user enters her ratings $r = (r_1, \ldots, r_{|\mathcal{S}|}) \in \mathbb{R}^{|\mathcal{S}|}$, her private variable x_0 *as well as* the disclosure $\ell = L(\mathcal{V}) \in \mathcal{L}^{|\mathcal{S}|}$. The program combines these quantities computing the obfuscated value $y = Y(r, x_0, \ell) \in \mathcal{Y}$, which the user subsequently reveals to the analyst.

Estimator. Finally, using the obfuscated output by the user, and the item profiles, the analyst constructs an estimator of the user's profile $\mathbf{x} \in \mathbb{R}^d$. Formally:

$$\widehat{\mathbf{x}} : \mathcal{Y} \times (\mathbb{R}^{(d+1)}_{-\mathbf{0}})^{|\mathcal{S}|} \to \mathbb{R}^d.$$

That is, given the item feature vectors $\mathcal{V} \subset \mathbb{R}^{d+1}_{-\mathbf{0}}$ and the corresponding obfuscated user feedback $y \in \mathcal{Y}$, it yields an estimate $\widehat{\mathbf{x}}(y, \mathcal{V})$ of the user's non-private feature vector $\mathbf{x}$. The estimator is a program executed by the analyst.

We refer to a triplet $\mathcal{R} = (L, Y, \widehat{\mathbf{x}})$ as a *learning protocol.* Note that the functional forms of all three of these components are known to both parties: *e.g.*, the analyst knows the obfuscation scheme Y. Both parties are *honest but curious*: they follow the protocol, but if at any step they can extract more information than what is intentionally revealed, they do so. All three mappings in protocol $\mathcal{R}$ can be randomized. In the following, we denote by $\mathrm{P}_{x,\mathcal{V}}$, $\mathrm{E}_{x,\mathcal{V}}$ the probability and expectation with respect to the noise in (7) as well as protocol randomization, given x, $\mathcal{V}$.

4.3 Privacy, Accuracy, and Disclosure Extent

Having formally specified a learning protocol $\mathcal{R} = (L, Y, \widehat{\mathbf{x}})$, we now define the three quality metrics we wish to attain, corresponding to the properties (a)-(c) of Section 4.1.

Privacy. We begin with our formalization of privacy:

[2] For technical reasons $\mathcal{L}$, and $\mathcal{Y}$ below, are in fact measurable spaces, which include of course $\mathbb{R}^k$, for some $k \in \mathbb{N}$.

Definition 1. *We say that $\mathcal{R} = (L, Y, \widehat{\mathbf{x}})$ is* privacy preserving *if the obfuscated output Y is independent of x_0. Formally, for all $\mathbf{x} \in \mathbb{R}^d$, $\mathcal{V} \subseteq \mathbb{R}^{(d+1)}_{-\mathbf{0}}$, and $A \subseteq \mathcal{Y}$,*

$$\mathrm{P}_{(-1,\mathbf{x}),\mathcal{V}}\big(Y(r,-1,\ell)\in A\big)=\mathrm{P}_{(+1,\mathbf{x}),\mathcal{V}}\big(Y(r,+1,\ell)\in A\big), \tag{8}$$

where $\ell = L(\mathcal{V})$ is the information disclosed from $\mathcal{V}$, and $r \in \mathbb{R}^{|\mathcal{S}|}$ is the vector of user ratings.

Intuitively, a learning protocol is privacy-preserving if its obfuscation scheme reveals nothing about the user's private variable: the distribution of the output Y does not depend statistically on x_0. Put differently, two users that have the same $\mathbf{x}$, but different x_0, output obfuscated values that are *computationally indistinguishable* [17].

Computational indistinguishability is a strong privacy property, as it implies a user's private variable is protected against *any* inference algorithm (and not just, e.g., the LSE (6)): in particular, no inference algorithm can estimate x_0 with probability better than 50% with access to y alone.

Accuracy. Our second definition determines a partial ordering among learning protocols w.r.t. their accuracy, as captured by the ℓ_2 loss of the estimation:

Definition 2. *We say that a learning protocol $\mathcal{R} = (L, Y, \widehat{\mathbf{x}})$ is more accurate than $\mathcal{R}' = (L', Y', \widehat{\mathbf{x}}')$ if, for all $\mathcal{V} \subseteq \mathbb{R}^{d+1}_{-\mathbf{0}}$,*

$$\sup_{\substack{x_0\in\{0,1\}\\ \mathbf{x}\in\mathbb{R}^d}} \mathrm{E}_{(x_0,\mathbf{x}),\mathcal{V}}\{\|\widehat{\mathbf{x}}(y,\mathcal{V})-\mathbf{x}\|_2^2\} \le \sup_{\substack{x_0\in\{0,1\}\\ \mathbf{x}\in\mathbb{R}^d}} \mathrm{E}_{(x_0,\mathbf{x}),\mathcal{V}}\{\|\widehat{\mathbf{x}}'(y',\mathcal{V})-\mathbf{x}\|_2^2\}\,,$$

where $y = Y(r, x_0, L(\mathcal{V}))$, $y' = Y'(r, x_0, L'(\mathcal{V}))$. Further, we say that it is strictly more accurate if the above inequality holds strictly for some $\mathcal{V} \subseteq \mathbb{R}^{d+1}_{-\mathbf{0}}$.

Note that the accuracy of $\mathcal{R}$ is determined by the ℓ_2 loss of the estimate $\widehat{\mathbf{x}}$ computed in a *worst-case* scenario, among all possible extended user profiles $x = (x_0, \mathbf{x})$.

This metric is natural. As we discuss in Section 6.2, it relates to the so-called A-optimality criterion [6]. It also has an additional compelling motivation. Recall that $\widehat{\mathbf{x}}$ is used to estimate the rating for a new item through the inner product (3). An estimator $\widehat{\mathbf{x}}$ minimizing the expected ℓ_2 loss also minimizes the mean square prediction error over a new item. This further motivates this accuracy metric, given that the analyst's goal is correct rating prediction.

To see this, assume that the extended user profile is estimated as $\widehat{x} = (\widehat{x}_0, \widehat{\mathbf{x}})$ for some $\mathbf{x}_0$ (for brevity we omit the dependence on $y, \mathcal{V}$). Recall that the analyst uses this profile to predict ratings for $v \notin \mathcal{V}$ using $\widehat{r} = \langle v, \widehat{x}\rangle$. The quality of such a prediction is often evaluated in terms of the mean square error (MSE):

$$\mathrm{MSE} = \mathrm{E}\{(r-\widehat{r})^2\} \overset{(7)}{=} \sigma^2 + \mathrm{E}\{\langle v, (x-\widehat{x})\rangle^2\}\,.$$

Assuming a random item vector v with diagonal covariance $\mathrm{E}(v_0^2) = c_0$, $\mathrm{E}(v_0\mathbf{v}) = 0$, $\mathrm{E}(\mathbf{v}\mathbf{v}^\mathsf{T}) = c\mathrm{I}$, we get

$$\mathrm{MSE} = \sigma^2 + c_0\mathrm{E}\{(x_0 - \widehat{x}_0)^2\} + c\mathrm{E}\big\{\|x - \widehat{x}\|_2^2\big\}\,.$$

Observe that the first term is independent of the estimation. Under a privacy-preserving protocol, the value for $\widehat{x}_0$ that minimizes the second term is 0.5, also independent of the estimation. The last term is precisely the ℓ_2 loss. Hence, minimizing the mean square error of the analyst's prediction is equivalent to minimizing the ℓ_2 loss of the estimator $\widehat{x}$. This directly motivates our accuracy definition.

Disclosure Extent. Finally, we define a partial ordering among learning protocols w.r.t. the amount of information revealed by their disclosures.

Definition 3. *We say that $\mathcal{R} = (L, R, \widehat{\mathbf{x}})$ discloses at least as much information as $\mathcal{R}' = (L', Y', \widehat{\mathbf{x}}')$ if there exists a measurable mapping $\varphi : \mathcal{L} \to \mathcal{L}'$ such that*

$$L' = \varphi \circ L$$

i.e., *$L'(v) = \varphi(L(v))$ for each $v \in \mathbb{R}^{d+1}_{-\mathbf{0}}$. We say that $\mathcal{R}$ and $\mathcal{R}'$ disclose the same amount of information if $L = \varphi \circ L'$ and $L' = \varphi' \circ L$ for some φ, φ'. Finally, we say that $\mathcal{R}$ discloses strictly more information than $\mathcal{R}'$ if $L' = \varphi \circ L$ for some φ but there exists no φ' such that $L = \varphi' \circ L'$.*

The above definition is again natural. Intuitively, a disclosure L carries at least as much information as L' if L' can be retrieved from L: the existence of the mapping φ implies that the user can recover L' from L by applying φ to the disclosure $L(\mathcal{V})$. Put differently, having a "black box" that computes L, the user can compute L' by feeding the output of this box to φ. If this is the case, then L is clearly at least as informative as L'.

5 An Optimal Protocol

In this section we prove that a simple learning protocol outlined in Algorithm 1, which we refer to as the *midpoint protocol* (MP), has remarkable optimality properties. The three components $(L, Y, \widehat{\mathbf{x}})$ of MP are as follows:

1. The analyst discloses the entry v_0 corresponding to the private user feature x_0, *i.e.*, $L\big((v_0, \mathbf{v})\big) \equiv v_0$ for all $(v_0, \mathbf{v}) \in \mathbb{R}^{d+1}_{-\mathbf{0}}$, and $\mathcal{L} \equiv \mathbb{R}$.
2. The user shifts each rating r_j by the contribution of her private feature. More specifically, the user reveals to the analyst the quantities:

   $$y_j = r_j - x_0 \cdot \ell_j = r_j - x_0 \cdot v_{j0}, \quad j \in \mathcal{S}.$$

   The user's obfuscated feedback is thus $Y(r, x_0, \ell) \equiv y$, where vector y's coordinates are the above quantities, i.e., $y = (y_1, \ldots, y_{|S|})$, and $\mathcal{Y} \equiv \mathbb{R}^{|\mathcal{S}|}$. Note that, by (7), for every $j \in \mathcal{S}$ the obfuscated feedback satisfies $y_j = \langle \mathbf{v}_j, \mathbf{x}\rangle + \varepsilon_j$, with ε_j the i.i.d. zero-mean noise variables.

Algorithm 1 Midpoint Protocol

```
Analyst's Parameters
$\mathcal{S} \subseteq [M]$, $\mathcal{V} = \{(v_{j0}, \mathbf{v}_j),\ j \in \mathcal{S}\} \subseteq \mathbb{R}^{d+1}_{-\mathbf{0}}$

User's Parameters
$x_0 \in \{-1, +1\}$, $r = (r_1, \ldots, r_{|\mathcal{S}|}) \in \mathbb{R}^{|\mathcal{S}|}$

DISCLOSURE: $\ell = L(\mathcal{V})$
$\ell_j = v_{j0}$, for all $j \in \mathcal{S}$

OBFUSCATION SCHEME: $y = Y(r, x_0, \ell)$
$y_j = r_j - x_0 \cdot \ell_j$, for all $j \in \mathcal{S}$

ESTIMATOR: $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}(y, \mathcal{V})$
Apply the minimax estimator $\widehat{\mathbf{x}}^*$ given by (9).
```

3. Finally, the analyst applies a *minimax* estimator on the obfuscated feedback. Let $\mathcal{X}$ be the set of all measurable mappings $\widehat{\mathbf{x}}$ estimating $\mathbf{x}$ given y and $\mathcal{V}$ (i.e., of the form $\widehat{\mathbf{x}} : \mathbb{R}^{|\mathcal{S}|} \times (\mathbb{R}^{(d+1)}_{-\mathbf{0}})^{|\mathcal{S}|} \to \mathbb{R}^d$). Estimator $\widehat{\mathbf{x}}^* \in \mathcal{X}$ is *minimax* if it minimizes the worst-case ℓ_2 loss, i.e.:

$$\sup_{x_0 \in \{0,1\}, \mathbf{x} \in \mathbb{R}^d} \mathrm{E}_{(x_0,\mathbf{x}),\mathcal{V}}\{\|\widehat{\mathbf{x}}^*(y,\mathcal{V}) - \mathbf{x}\|_2^2\} = \inf_{\widehat{\mathbf{x}} \in \mathcal{X}} \sup_{x_0 \in \{0,1\}, \mathbf{x} \in \mathcal{X}} \mathrm{E}_{(x_0,\mathbf{x}),\mathcal{V}}\{\|\widehat{\mathbf{x}}(y,\mathcal{V}) - \mathbf{x}\|_2^2\}. \tag{9}$$

The following theorem summarizes the midpoint protocol's remarkable properties:

Theorem 1. *Under the linear model* (7)*:*

1. MP is privacy preserving.

2. No privacy preserving protocol is strictly more accurate than MP.

3. Any privacy preserving protocol that does not disclose at least as much information as MP is strictly less accurate.

We prove the theorem below. Its second and third statement establish formally the optimality of the midpoint protocol. Intuitively, the second statement implies that the midpoint protocol has *maximal accuracy.* No privacy preserving protocol achieves better accuracy: surprisingly, this is true even among schemes that *disclose strictly more information than the midpoint protocol.* As such, the second statement of the theorem imples there is no reason to disclose more than v_{j0} for each item $j \in \mathcal{S}$.

The third statement implies that the midpoint protocol engages in a *minimal disclosure*: to achieve maximal accuracy, a learning protocol *must disclose at least* v_{j0}, $j \in \mathcal{S}$. In fact, our proof shows that the gap between the accuracy of MP and a protocol not disclosing biases is infinite, for certain $\mathcal{V}$. We note that the disclosure in MP is intuitively appealing: an analyst need only disclose the gap between average ratings across the two types (e.g., males and females, conservatives and liberals, etc.) to enable protection of x_0.

In general, the minimax estimator $\widehat{\mathbf{x}}^*$ depends on the distribution followed by the noise variables in (7). For arbitrary distributions, a minimax estimator that can be computed in a closed form (rather than as the limit of a sequence of estimators) may not be known. General conditions for the existence of such estimators can be found, e.g., in Strasser [41]. In the case of Gaussian noise, the minimax estimator coincides with the least squares estimator (see, e.g., Lehman and Casella [26, Thm. 1.15, Chap. 5]), i.e.,

$$\widehat{\mathbf{x}}^*(y,\mathcal{V})=\arg\min_{\mathbf{x}\in\mathbb{R}^d}\Big\{\textstyle\sum_{j=1}^{|\mathcal{S}|}\big(y_j-\langle\mathbf{v}_j,\mathbf{x}\rangle\big)^2\Big\}. \tag{10}$$

The minimization in (10) is a linear regression, and $\widehat{\mathbf{x}}^*$ has the following closed form:

$$\widehat{\mathbf{x}}^*(y,\mathcal{V}) = \big(\textstyle\sum_{j\in\mathcal{S}}\mathbf{v}_j\mathbf{v}_j^T\big)^{-1}\cdot\big(\textstyle\sum_{j\in\mathcal{S}}y_j\mathbf{v}_j\big). \tag{11}$$

The accuracy of this estimator can also be computed in a closed form. Using, (7), (11), and the definition of y, it can easily be shown that, for all $\mathbf{x}\in\mathbb{R}^d$,

$$\mathrm{E}_{(x_0,\mathbf{x}),\mathcal{V}}\{\|\widehat{\mathbf{x}}^*(y,\mathcal{V})-\mathbf{x}\|_2^2\}=\sigma^2\mathrm{tr}\big[\big(\textstyle\sum_{j\in\mathcal{S}}\mathbf{v}_j\mathbf{v}_j^T\big)^{-1}\big], \tag{12}$$

where σ^2 the noise variance in (7) and $\mathrm{tr}(\cdot)$ the trace.

5.1 Proof of Theorem 1

Privacy. To see that Thm. 1.1 holds, observe that the user releases $y_j = r_j - v_{0j}x_0 \overset{(7)}{=} \langle\mathbf{v}_j,\mathbf{x}\rangle+\varepsilon_j$, for each $j\in\mathcal{S}$. The distribution of y_j thus does not depend on x_0, so the midpoint protocol is clearly privacy preserving.

Maximal Accuracy. We prove Theorem 1.2 by contradiction; in particular, we show that a protocol that is strictly more accurate can be used to construct an estimator that has lower worst-case ℓ_2 loss than the minimax estimator.

Suppose that there exists a privacy preserving protocol $\mathcal{R}' = (L', Y', \widehat{\mathbf{x}}')$ that is strictly more accurate than the midpoint protocol $\mathcal{R} = (L, Y, \widehat{\mathbf{x}})$. Let $\ell = L(\mathcal{V}), \ell' = L'(\mathcal{V})$ be the disclosures under the two protocols, and $y = Y(r, x_0, \ell)$, $y' = Y'(r, x_0, \ell')$ the obfuscated values. Recall that

$$\ell_j = v_{j0},\ \text{and}\ y_j = r_j - x_0v_{0j} = \langle\mathbf{v}_j,\mathbf{x}\rangle+\varepsilon_j,\quad j\in\mathcal{S}.$$

We will use L', Y' and $\widehat{\mathbf{x}}'$ to construct an estimator $\widehat{\mathbf{x}}''$ that has a lower ℓ_2 loss than the least squares estimator $\widehat{\mathbf{x}}$ over y and $\mathcal{V}$. First, apply Y' to $y+\ell$, assuming that the private variable is $x_0=+1$, using the disclosed information ℓ'. That is: $y'' = Y'(y+\ell,+1,\ell')$. Second, apply the estimator $\widehat{\mathbf{x}}'$ to this newly obfuscated output y'', i.e.: $\widehat{\mathbf{x}}'(y'',\mathcal{V})$ Combining these two the estimator $\widehat{\mathbf{x}}''$ is given by

$$\widehat{\mathbf{x}}''(y,\mathcal{V}) = \widehat{\mathbf{x}}'\left(Y'\left(y+\ell,+1,L'(\mathcal{V})\right),\mathcal{V}\right)$$

Under this construction, the random variables y'', y' are identically distributed. This is obvious if $x_0 = +1$; indeed, in this case $y'' = y'$. On the other hand, since $\mathcal{R}'$ is privacy preserving, by (8):

$$Y'(y+\ell,+1,\ell') \overset{\mathrm{d}}{=} Y'(y-\ell,-1,\ell'), \tag{13}$$

i.e., the two variables are equal in distribution.

This implies that $\widehat{\mathbf{x}}''(y,\mathcal{V})$ is identically distributed as $\widehat{\mathbf{x}}'(y',\mathcal{V})$. On the other hand, $\mathcal{R}'$ is strictly more accurate than $\mathcal{R}$; hence, there exists a $\mathcal{V}$ such that

$$\begin{aligned}\sup_x \mathrm{E}\{\|\widehat{\mathbf{x}}(y,\mathcal{V})-\mathbf{x}\|_2^2\} &> \sup_x \mathrm{E}\{\|\widehat{\mathbf{x}}'(y',\mathcal{V})-\mathbf{x}\|_2^2\} \\ &= \sup_x \mathrm{E}\{\|\widehat{\mathbf{x}}''(y,\mathcal{V})-\mathbf{x}\|_2^2\},\end{aligned}$$

a contradiction.

Minimal Disclosure. Consider a privacy preserving learning protocol $\mathcal{R}' = (L', Y', \widehat{\mathbf{x}}')$ that does not disclose at least as much information as the midpoint protocol $\mathcal{R} = (L, Y, \widehat{\mathbf{x}})$. Consider a setup where $|\mathcal{S}| = d$, the dimension of the feature profiles. Assume also that $\mathcal{V}$ is such that the matrix $V = [\mathbf{v}_j]_{j\in\mathcal{S}} \in \mathbb{R}^{d\times d}$ is invertible, and denote by $\ell = L(\mathcal{V}) \in \mathbb{R}^d$ the vector with coordinates v_{j0}, $j \in \mathcal{S}$.

For any $x_0 \in \{+1,-1\}$, $s \in \mathbb{R}^d$, and $\ell' \in (\mathcal{L}')^d$, let $Z_{x_0}(s,\ell') \in \mathcal{Y}'$ be a random variable with distribution given by $Z_{x_0}(s,\ell') \overset{\mathrm{d}}{=} Y'(s+\varepsilon, x_0, \ell')$, where $\varepsilon \in \mathbb{R}^d$ a vector of *i.i.d.* coordinates sampled from the same distribution as the noise variables ε_j, $j \in \mathcal{S}$. Put differently, $Z_{x_0}(s,\ell')$ is distributed as the output of obfuscation Y' when $r - \varepsilon = V\mathbf{x} + x_0\ell = s \in \mathbb{R}^d$, $L'(\mathcal{V}) = \ell'$, and the gender is x_0. The following then holds.

Lemma 1. *If $V \in \mathbb{R}^{d\times d}$ is invertible then, for all $s \in \mathbb{R}^d$, $\ell = L(\mathcal{V})$, and $\ell' = L'(\mathcal{V})$, $Z_+(s,\ell') \overset{\mathrm{d}}{=} Z_-(s-2\ell,\ell')$.*

Proof. By Eq. (13), for all $\mathbf{x} \in \mathbb{R}^d$,

$$Y'(V\mathbf{x}+\ell+\varepsilon,+1,\ell') \overset{\mathrm{d}}{=} Y'(V\mathbf{x}-\ell+\varepsilon,-1,\ell')\,.$$

The claim follows by taking $\mathbf{x} = V^{-1}(s-\ell)$. □

As $\mathcal{R}'$ does not disclose as much information as the midpoint protocol, by definition, there is no map φ such that $L(v) = \varphi(L'(v))$ for all $v = (v_0, \mathbf{v}) \in \mathbb{R}^{d+1}_{-\mathbf{0}}$. Hence, there exist extended profiles $v, v' \in \mathbb{R}^{d+1}_{-\mathbf{0}}$ such that $v_0 \neq v'_0$ and yet $L'(v) = L'(v')$. As both $v = (v_0,\mathbf{v}), v' = (v'_0,\mathbf{v}')$ belong to $\mathbb{R}^{d+1}_{-\mathbf{0}}$, the supports of $\mathbf{v}, \mathbf{v}'$ are non-empty. We consider the following two cases:

Case 1. The supports of $\mathbf{v}, \mathbf{v}'$ intersect, i.e., there exists a $k \in [d]$ such that $v_k \neq 0$ and $v'_k \neq 0$. In this case, consider a scenario in which $\mathcal{V} = \{v\} \cup \bigcup_{1\le l\le d, l\neq k,}\{e_l\}$, where $e_l \in \mathbb{R}^{d+1}_{-\mathbf{0}}$ a vector whose l-th coordinate is 1 and all other coordinates are zero. Clearly, $|\mathcal{S}| = |\mathcal{V}| = d$, and $V = [\mathbf{v}_i]_{i\in[d]}$ is invertible. Let $\ell^* = L'(\mathcal{V})$. By Lemma 1, for all $s \in \mathbb{R}$,

$$Z_+(s + 2v_0\mathbf{e}_1, \ell^*) \overset{\mathrm{d}}{=} Z_-(s, \ell^*)\,, \tag{14}$$

where $\mathbf{e}_1 \in \mathbb{R}^d$ is 1 at coordinate 1 and 0 everywhere else. Similarly, in a scenario in which $\mathcal{V}' = \{v'\} \cup \bigcup_{1\leq l\leq d, l\neq k,}\{e_l\}$, V is again invertible. Crucially $L'(\mathcal{V}') = L(\mathcal{V}) = \ell^*$, so again by Lemma 1:

$$Z_+(s + 2v'_0\mathbf{e}_1, \ell^*) \overset{\mathrm{d}}{=} Z_-(s, \ell^*)\,, \tag{15}$$

for all $s \in \mathbb{R}^d$. Equations (14),(15) imply that, for all $s \in \mathbb{R}^d$:

$$Z_+(s + \xi\mathbf{e}_1, \ell^*) \overset{\mathrm{d}}{=} Z_+(s, \ell^*) \tag{16}$$

where $\xi \equiv 2(v_0 - v'_0)$. In other words, the obfuscation is *periodic* with respect to the direction $\mathbf{e}_1$.

Observe that for any $x \in \{-1, +1\} \times \mathbb{R}^d$ and any $M \in \mathbb{R}_+$, we can construct a $x' \in \{-1, +1\} \times \mathbb{R}^d$ and a $K \in \mathbb{N}$ such that (a) x,x' differ only at coordinate $k \in \{1, 2, \ldots, d\}$, (b) $\langle v, x - x'\rangle = K\xi$, and (c) $\|\mathbf{x} - \mathbf{x}'\|_2 \geq M$. To see this, let K be a large enough integer such that $\frac{K|\xi|}{|v_k|} > M$. Taking, $x'_k = x_k + K\xi/v_k$, and $x'_l = x_l$ for all other l in $\{0, 1, \ldots, d\}$ yields a x' that satisfies the desired properties (a) and (b).

Suppose that the learning protocol $\mathcal{R}$ is applied to $\mathcal{V} = \{v\} \cup \bigcup_{1\leq l\leq d, \neq k}\{e_l\}$ for a user with $x_0 = +1$. Fix a large $M > 0$. For each x and x' constructed as above, by (16), the obfuscated values generated by Y' have an identical distribution. Hence, irrespectively of how the estimator $\widehat{\mathbf{x}}'$ is implemented, either $\mathrm{E}_{x,\mathcal{V}}\{\|\widehat{\mathbf{x}}'(y', \mathcal{V}) - \mathbf{x}\|_2^2\}$ or $\mathrm{E}_{x',\mathcal{V}}\{\|\widehat{\mathbf{x}}'(y', \mathcal{V}) - \mathbf{x}'\|_2^2\}$ must be $\Omega(M^2)$ which, in turn, implies that $\sup_{x\in\{\pm1\}\times\mathbb{R}^d} \mathrm{E}_{x,\mathcal{V}}\{\|\widehat{\mathbf{x}}'(y', \mathcal{V}) - \mathbf{x}\|_2^2 = \infty$.

Note that, since $\mathbf{v}_j, j \in \mathcal{S}$, are linearly independent, the matrix $\sum_{j\in\mathcal{S}} \mathbf{v}_j\mathbf{v}_j^T$ is positive definite and thus invertible. Hence, in contrast to the above setting, the loss (12) of MP in the case of Gaussian noise is finite.

Case 2. The supports of $\mathbf{v}, \mathbf{v}'$ are disjoint. In this case v, v' are linearly independent and, in particular, there exist $1 \leq k, k' \leq d$, $k \neq k'$, such that $v_k \neq 0$, $v_{k'} = 0$ while $v'_k = 0$, $v'_{k'} \neq 0$. Let $\mathcal{V} = \{v\} \cup \{v'\} \bigcup_{1\leq l\leq d, l\neq k, l\neq k'}\{e_l\}$. Then, $|\mathcal{V}| = d$ and the matrix $V = [\mathbf{v}_j]_{j\in\mathcal{S}}$ is again invertible. By swapping the positions of $\mathbf{v}$ and $\mathbf{v}'$ in matrix V we can show using a similar argument as in Case 1 that for all $s \in \mathbb{R}^d$:

$$Z_+(s + \xi(\mathbf{e}_1 - \mathbf{e}_2), \ell^*) \overset{\mathrm{d}}{=} Z_+(s, \ell^*) \tag{17}$$

where $\xi \equiv 2(v_0 - v'_0)$ and $\ell^* = L(\mathcal{V})$. *I.e.*, Z_+ is periodic in the direction $\mathbf{e}_1 - \mathbf{e_2}$. Moreover, for any $x \in \{-1, +1\} \times \mathbb{R}^d$ and any $M \in \mathbb{R}_+$, we can similarly construct a $x' \in \{-1, +1\} \times \mathbb{R}^d$ and a $K \in \mathbb{N}$ such that (a) x,x' differ only at coordinates $k, k' \in \{1, 2, \ldots, d\}$, and (b) $\langle v, x - x'\rangle = -\langle v', x - x'\rangle = K\xi$, and (c) $\|\mathbf{x} - \mathbf{x}'\|_2 \geq M$: the construction adds $K\xi/v_k$ at the k-th coordinate and subtracts $K\xi/v'_{k'}$ from the k'-th coordinate, where $K > M \max(v_k, v'_{k'})/\xi$. A similar argument as in Case 1 can be used to show again that the estimator $\widehat{\mathbf{x}}'$ cannot disambiguate between x, x' over $\mathcal{V}$, yielding the theorem. □

Algorithm 2 Midpoint Protocol with Sub-Sampling

```
Analyst's Parameters
$\mathcal{S} \subseteq [M]$, $\mathcal{V} = \{(v_{j0}, \mathbf{v}_j),\ j \in \mathcal{S}\} \subseteq \mathbb{R}^{d+1}_{-\mathbf{0}}$
$p = \{(p_1^{x_0}, \ldots, p_{|\mathcal{S}|}^{x_0}), x_0 \in \{-,+\}\} \subseteq ([0,1] \times [0,1])^{|\mathcal{S}|}$

User's Parameters
$x_0 \in \{-1,+1\}$, $\mathcal{S}_0 \subseteq \mathcal{S}$, $r = \{r_j, j \in \mathcal{S}_0\} \in \mathbb{R}^{|\mathcal{S}_0|}$

DISCLOSURE: $\ell = L(\mathcal{V}, p)$
$\rho_j = p_j^- / p_j^+$, for all $j \in \mathcal{S}$
$\ell_j = (v_{j0}, \rho_j)$, for all $j \in \mathcal{S}$

OBFUSCATION SCHEME: $\substack{\mathcal{S}_R = \mathcal{S}_R(\mathcal{S}_0, x_0, \ell) \\ y = Y(r_{\mathcal{S}_\mathcal{R}}, x_0, \ell)}$
$S_R = \emptyset$, $y = \emptyset$
for all $j \in \mathcal{S}$ do
  if $j \in \mathcal{S}_0$ then
    $b_j \sim \mathrm{Bern}\big(\min\big(1, (\rho_j)^{x_0}\big)\big)$
    if $b_j = 1$ then
      $\mathcal{S}_R = \mathcal{S}_R \cup \{j\}$
      $y = y \cup \{r_j - x_0 v_{j0}\}$
    end if
  end if
end for

ESTIMATOR: $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}(y, (\mathcal{S}_R, y), \mathcal{V})$
Solve $\widehat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathbb{R}^d} \Big\{\sum_{j \in \mathcal{S}_R} \big(y_j - \langle \mathbf{v}_j, \mathbf{x} \rangle\big)^2\Big\}$
```

6 Extensions

We have up until now assumed that the analyst solicits ratings for a set of items $\mathcal{S} \subseteq [M]$, determined by the analyst before the user reveals her feedback. In what follows, we discuss how our analysis can be extended in the case where the user only provides ratings for a subset of these items. We also discuss how the analyst should select $\mathcal{S}$, how a user can repeatedly interact with the analyst, and, finally, how to deal with multiple binary and categorical features.

6.1 Partial Feedback

There are cases of interest where a user may not be able to generate a rating for all items in $\mathcal{S}$. This is especially true when the user needs to spend a non-negligible effort to determine her preferences (examples include rating a feature-length movie, a restaurant, or a book). In these cases, it makes sense to assume that a user may readily provide ratings for only a set $\mathcal{S}_0 \subseteq \mathcal{S}$ (e.g., the movies she has already watched, or the restaurants she has already dined in, etc.).

Our analysis up until now applies when the user rates an arbitrary set $\mathcal{S}$

selected by the analyst. As such, it does not readily apply to this case: the set of items $\mathcal{S}_0$ a user rates may depend on the private feature x_0 (e.g., some movies may be more likely to be viewed by men or liberals). In this case, x_0 would be be inferable not only from the ratings she gives, but also *from which items she has rated.*

In this section, we describe how to modify the midpoint protocol to deal with this issue. Intuitively, to ensure her privacy, rather than reporting obfuscated ratings for *all* items she rated (i.e., set $\mathcal{S}_0$), the user can reveal ratings *only for a subset* $\mathcal{S}_R$ *of* $\mathcal{S}_0$. This *sub-sampling* of $\mathcal{S}_0$ can be done so that $\mathcal{S}_R$ has a distribution that is *independent of* x_0, even though $\mathcal{S}_0$ does not. Moreover, to ensure a high estimation accuracy, the user ought to ensure $\mathcal{S}_R$ is as large as possible, subject to the constraint $\mathcal{S}_R \subseteq \mathcal{S}_0$.

Model. Before we present the modified protocol, we describe our assumption on how $\mathcal{S}_0$ is generated. For each $j \in [M]$, denote by p_j^+, p_j^- the probabilities that a user with private feature $+1$ or -1, respectively, has rated item $j \in \mathcal{M}$. Observe that, just like the extended profiles v_j, this information can be extracted from a dataset comprising ratings by non privacy-conscious users. Let $p = [(p_j^+, p_j^-)]_{j\in\mathcal{S}} \in ([0,1] \times [0,1))^{|\mathcal{S}|}$ be the vector of pairs of probabilities.

We assume that the privacy-conscious user the has rated items in the set $S_0 \subseteq \mathcal{S}$, whose distribution is given by the product form[3]:

$$\mathrm{P}_{x,\mathcal{V},p}(\mathcal{S}_0 = \mathcal{A}) = \prod_{j\in\mathcal{S}} p_j^{x_0} \prod_{j\in\mathcal{S}\setminus\mathcal{A}} (1 - p_j^{x_0}), \text{ for all } \mathcal{A} \subseteq \mathcal{S}. \tag{18}$$

Put differently, items $j \in \mathcal{S}$ are rated independently, each with a probability $p_j^{x_0}$. Conditioned on $\mathcal{S}_0$, we assume that the user's ratings r_j, $j \in \mathcal{S}_0$, follow the linear model (7) with Gaussian noise. Note that the distribution of $\mathcal{S}_0$ depends on x_0: e.g., items j for which $p_j^+ > p_j^-$ are more likely to be rated when $x_0 = +1$.

Midpoint Protocol with Sub-Sampling. We now present a modification of the midpoint protocol, which we refer to as the *midpoint protocol with sub-sampling* (MPSS). MPSS is summarized in Algorithm 2. First, along with the disclosure of the biases $v_{j0}, j \in \mathcal{S}$, the analyst also discloses the ratios $\rho_j \equiv p_j^- / p_j^+$, for $j \in \mathcal{S}$. Having access to this information, the user sub-samples items from $\mathcal{S}_0$; each item $j \in S_0$ is included in the revealed set $\mathcal{S}_R$ independently with probability:

$$\mathrm{P}_{x,\mathcal{V},p}(i \in \mathcal{S}_R \mid j \in S_0) = \min\left(1, (\rho_j)^{x_0}\right) . \tag{19}$$

Having constructed $\mathcal{S}_R$, the user reveals ratings for $j \in S_R$ after subtracting $x_0 v_{j0}$, as in MP. Finally, the analyst estimates $\widehat{\mathbf{x}}$ through a least squares estimation over the obfuscated feedback, as in MP in the case of Gaussian noise.

To gain some intuition behind the selection of the set $\mathcal{S}_R$, observe by (18) and (19) that, for any $j \in \mathcal{S}$,

$$\mathrm{P}_{x,\mathcal{V},p}(j \in \mathcal{S}_R) = p_j^{x_0} \min\left(1, (p_j^-/p_j^+)^{x_0}\right) = \min(p_j^+, p_j^-). \tag{20}$$

[3] Here, we slightly abuse notation, e.g., denoting with $p_j^{x_0}$ the parameter p_j^+ when $x_0 = +1$.

This immediately implies that MPSS is privacy preserving: both the distribution of $\mathcal{S}_R$ and of the obfuscated ratings y do not depend on x_0. In fact, it is easy to see that since $\mathcal{S}_R \subseteq \mathcal{S}_0$ *any privacy preserving protocol must satisfy* $\mathrm{P}_{x,\mathcal{V},p}(j \in \mathcal{S}_R) \leq \min(p_j^+, p_j^-)$: indeed, if for example $p_j^+ < p_j^-$, then a user rating j with probability higher than p_j^+ must have $x_0 = -1$ (see Lemma 2 in the appendix for a formal proof of this statement). As such, MPSS reveals ratings for a set $\mathcal{S}_R$ of *maximal size*, in expectation.

This intuition can be used to establish the optimality of MPSS among a wide class of learning protocols, under (7) (with Gaussian noise) and (18). We can again show that it attains optimal accuracy. Moreover, it also involves a minimal disclosure: a protocol that does not reveal the ratios ρ_j, $j \in \mathcal{S}$, necessarily rates strictly fewer items than MPSS, in expectation. We provide a formal proof of these statements, as well as a definition of the class of protocols we consider, in Appendix A.

6.2 Item Set Selection

Theorem 1 implies that the analyst *cannot* improve the prediction of the private variable x_0 through its choice of $\mathcal{S}$, under the midpoint protocol. In fact, the same is true under any privacy-preserving learning protocol: irrespectively of the analyst's choice for $\mathcal{S}$, the obfuscated feedback y will be statistically independent of x_0.

The analyst can however strategically select $\mathcal{S}$ to effect the accuracy of the estimate of the non-private profile $\mathbf{x}$. Indeed, the analyst should attempt to select a set $\mathcal{S}$ that maximizes the accuracy of the estimator $\widehat{\mathbf{x}}$. In settings where least squares estimator is minimax (e.g., when noise is Gaussian), there are well-known techniques for addressing this problem. Eq. (12) implies that it is natural to select $\mathcal{S}$ by solving

$$\begin{aligned} \text{Maximize:} \quad & F(\mathcal{S}) = -\mathrm{tr}\big[\big(\textstyle\sum_{j\in\mathcal{S}} \mathbf{v}_j \mathbf{v}_j^T\big)^{-1}\big] \\ \text{subject to:} \quad & |\mathcal{S}| \leq B, \mathcal{S} \subseteq [M], \end{aligned} \tag{21}$$

where B is the number of items for which the analyst solicits feedback. The optimization problem (21) is NP-hard, and has been extensively studied in the context of experimental design (see, e.g., Section 7.5 of [6]). The objective function F is commonly referred to as the *A-optimality criterion.* Convex relaxations of (21) exist when $\mathcal{S}$ is a multiset, i.e., when items with the same profile can be presented to the user multiple times, each generating an i.i.d. response [6]. When such repetition is not possible, constant approximation algorithms can be constructed based on the fact that F is increasing and submodular (see, e.g., [16]). In particular, given any set $\mathcal{S}^* \subset [M]$ of items whose profiles are linearly independent, there exists a polynomial time algorithm for maximizing $F(\mathcal{S} \cup \mathcal{S}^*) - F(\mathcal{S}^*)$ subject to $|\mathcal{S}| \leq B$ within a $1 - \frac{1}{e}$ approximation factor [32].

6.3 Repeated Interactions

Our analysis (and, in particular, the optimality of our protocols) persists even when the user repeatedly interacts with the analyst. In particular, the user may return to the service multiple times, each time asked to rate a different set of items $\mathcal{S}^{(k)} \setminus [M]$, $k \geq 1$. The selection of the set $\mathcal{S}^{(k)}$ could be *adaptive*, i.e., depend on the obfuscated feedback the user has revealed up until the $k-1$-th time. For each k, the analyst again would apply MP (or MPSS), again disclosing the same information for each $\mathcal{S}^{(k)}$, the only difference being that the estimator $\widehat{\mathbf{x}}$ would be applied to *all* revealed obfuscated ratings $y^{(1)}, ..., y^{(k)}$. This repeated interaction is still perfectly private: the joint distribution of the obfuscated outputs $y^{(k)}$ does not depend on x_0. Moreover, each estimation remains maximally accurate at each interaction, while each disclosure is again minimal.

6.4 Categorical Features

We discuss below how to express categorical features as multiple binary features through binarization, and illustrate how to incorporate both cases in our analysis. The standard approach to incorporating a categorical feature $x_0 \in \{1, 2, \ldots, K\}$ in matrix factorization is through category-specific biases (see, e.g., [24]), i.e., (7) is replaced by

$$r = \langle \mathbf{x}, \mathbf{v}_j \rangle + b_j^{x_0} + \varepsilon_j, \quad j \in [M] \tag{22}$$

where $b_j^k \in \mathbb{R}$, $k \in [K]$ are *category-dependent* biases. Consider a representation of $x_0 \in [K]$ as a binary vector $\mathbf{x}_0 \in \{-1, +1\}^K$ whose x_0-th coordinate is $+1$, and all other coordinates are -1. I.e., the coordinate $\mathbf{x}_{0k}$ at $k \in [K]$ is given $+1$ if $k = x0$ and -1 o.w. For $k \in [K]$, let $\mathbf{b}_{jk} \equiv b_j^k/2$ and define $\mu_j \equiv \sum_{k \in [K]} b_j^k/2$. Then, observe that (22) is equivalent to

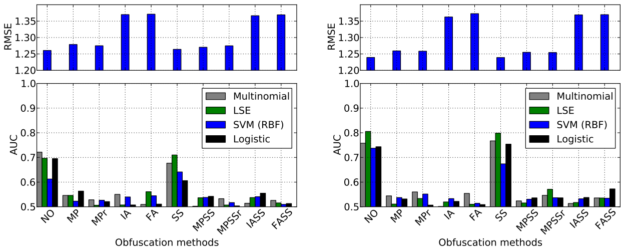
$$\begin{aligned} r &= \langle \mathbf{x}, \mathbf{v}_j \rangle + \sum_{k \in [K]} \mathbf{x}_{0k} \mathbf{b}_{jk} + \mu_j + \varepsilon_j, \\ &= \langle \mathbf{x}', \mathbf{v}'_j \rangle + \sum_{k \in [K]} \mathbf{x}_{0k} \mathbf{b}_{jk} + \varepsilon_j, \quad j \in [M], \end{aligned} \tag{23}$$

where $\mathbf{x}' = (\mathbf{x}, 1) \in \mathbb{R}^{d+1}$ and $\mathbf{v}'_j = (\mathbf{v}_j, \mu_j) \in \mathbb{R}^{d+1}$.

Hence, a categorical feature can be incorporated in our analysis as follows. First, given a dataset of ratings by non-privacy conscious users that reveal their categorical feature $x_0 \in [K]$, the analyst first “binarizes” this feature, constructing a vector $\mathbf{x}_0 \in \{-1, 1\}^K$ for each user. It then performs matrix factorization on the ratings using (23), learning vectors $\mathbf{v}'_j \in \mathbb{R}^{d+1}$, and biases $\mathbf{b}_j = (\mathbf{b}_{jk})_{k \in [K]} \in \mathbb{R}^K$. A privacy-conscious user subsequently interacts with the analyst using the standard scheme as follows. The analyst discloses the biases $\mathbf{b}_j$ for each $j \in \mathcal{S}$, and the user reveals $y_j = r_j - \sum_{k \in [K]} \mathbf{x}_{0k} \mathbf{b}_{jk}$, $j \in \mathcal{S}$, where $\mathbf{x}_{0k}$ is her binarized categorical feature. Finally, the analyst infers $\mathbf{x}'$ through linear regression over the pairs $(y_j, \mathbf{v}'_j)$, $j \in \mathcal{S}$.

Dataset	Private feature	Users	Items	Ratings
PTV	All	365	50	18K
	Gender (F:M)	2.7:1	-	2.7:1
	Politics (R:D)	1:1.4	-	1:1.4
Movielens	All	6K	3K	1M
	Gender (F:M)	1:2.5	-	1:3
	Age (Y:A)	1:1.3	-	1:1.6
Flixster	All	26K	9921	5.6M
	Gender (F:M)	1.7:1	-	1.5:1

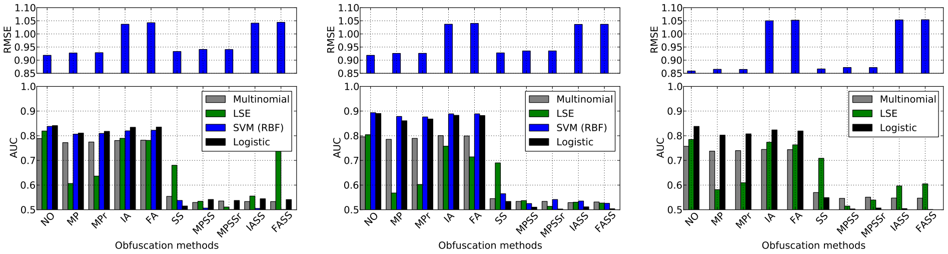
Figure 3: Statistics of the datasets used for evaluation. The ratios represent the class skew Females:Males (F:M) for gender, Young:Adult(Y:A) for age and Republican:Democrats (R:D) for political affiliation.



(a) PTV - Gender

(b) PTV - Politics

Figure 4: Privacy risk and prediction accuracy on PTV, obtained using four classifiers and obfuscation schemes (NO-No obfuscation, MP - Midpoint Protocol, r - Rounding, IA - Item Average, FA - Feature Average, SS - Sub-Sampling). The proposed protocol (MP) is robust to privacy attacks with hardly any loss in predictive power.



(a) Movielens - Gender

(b) Movielens - Age

(c) Flixster - Gender

Figure 5: Privacy risk and prediction accuracy on Movielens and Flixster (sparse datasets), obtained using four classifiers and obfuscation schemes (NO-No obfuscation, MP - Midpoint Protocol, r - Rounding, IA - Item Average, FA - Feature Average, SS - Sub-Sampling). The proposed protocol (MPSS) is robust to privacy attacks without harming predictive power.

7 Evaluation

In this section we evaluate our protocols on real-world datasets. Our experiments confirm that MP and MPSS are indeed able to protect the privacy of users against inference algorithms, including non-linear algorithms, with little impact on prediction accuracy.

7.1 Experimental Setup

We study two types of datasets, sparse and dense. In sparse datasets, the set of items a user rates is often correlated with the private feature. This does not happen in dense datasets because all users rate all (or nearly all) items.

Datasets. We evaluate our methods on three datasets: Politics-and-TV, Movielens and Flixster. Politics-and-TV (PTV) [38] is a ratings dataset that includes 5-star ratings of users to 50 TV-shows and, in addition, each user's political affiliation (Democrat or Republican) and gender. To make it dense, we consider only users that rate over 40 items, resulting in 365 users; 280 provide ratings to all 50 TV shows. Movielens[4] and Flixster[5] [20] are movie recommender systems in which users rate movies from a catalog of thousands of movies. Both Movielens and Flixster datasets include user gender. Movielens also includes age groups; we categorize users as *young adults* (ages 18–35), or *adults* (ages 35–65). We preprocessed Movielens and Flixster to consider only users with at least 20 ratings, and items that were rated by at least 20 users. Table 3 summarizes the statistics of these three datasets.

Methodology. Throughout the evaluation, we seek to quantify the privacy risk to a user as well as the impact of obfuscation on the prediction accuracy. To this end, we perform 10-fold cross-validation as follows. We split the users in each dataset into 10 folds. We use 9 folds as a *training set* (serving the purpose of a dataset of non privacy-conscious users in Figure 2) and 1 fold as a *test set* (whose users we treat as privacy-conscious).

We use the training set to (a) compute extended profiles for each item by performing matrix factorization, (b) empirically estimate the probabilities p^+, p^- for each item, and (c) train multiple classifiers, to be used to infer the private features. We describe the details of our MF implementation and the classifiers we use below.

We split the ratings of each user in the test set into two sets by randomly selecting 70% of the ratings as the first set, and the remaining 30% as the second set. We obfuscate the ratings in the first set using MP, MPSS, and several baselines as described in detail below. The obfuscated ratings are given as input to our classifiers to infer the user's private feature. We further estimate a user's extended profile using the LSE method described in Section 3.3, and use this profile (including both $\mathbf{x}$ and the inferred x_0) to predict her ratings on the second set. For each obfuscation scheme and classification method, we measure the *privacy risk* of the inference through these classifiers using the area under the curve (AUC) metric [18, 19]. Moreover, for each obfuscation scheme, we measure the *prediction accuracy* through the root mean square error (RMSE) of the predicted ratings. We cross-validate our results by computing the AUC and RMSE 10 times, each time with a different fold as a test set, and reporting average values. We note that the AUC ranges from 0.5 (perfect privacy) to 1 (no privacy).

Matrix Factorization. We use 20 iterations of stochastic gradient descent [24] to perform MF on each training set. For each item, feature biases v_{j0} were computed as the half distance between the average item ratings per each private

[4] `http://www.grouplens.org/node/73`

[5] `http://www.sfu.ca/~sja25/datasets/`

feature value. The remaining features $\mathbf{v}_j$ were computed through matrix factorization. We computed optimal regularization parameters and the dimension $d = 20$ through an additional 10-fold cross validation.

Privacy Risk Assessment. We apply several standard classification methods to infer the private feature from the training ratings, namely Multinomial Naïve Bayes [44], Logistic Regression (LR), non-linear Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel, as well as the LSE (6). The input to the LR, NB and SVM methods comprises the ratings of all items provided by the user as well as zeros for movies not rated, while LSE operates only on the ratings that the user provides. As SVM scales quadraticaly with the number of users, we could not execute it on our largest dataset (Flixster, c.f. Table 3).

Obfuscation Schemes. When using MP, the obfuscated rating may not be an integer value, and may even be outside of the range of rating values which is expected by a recommender system. Therefore, we consider a variation of MP that rounds the rating value to an integer in the range $[1, 5]$. Given a non-integer obfuscated rating r, which is between two integers $k = \lfloor r \rfloor$ and $k + 1$, we perform rounding by assigning the rating k with probability $r - k$ and the rating $k + 1$ with probability $1 - (r - k)$, which on expectation gives the desired rating r. Ratings higher than 5 and those lower than 1 are truncated to 5 and 1, respectively. We refer to this process as *rounding*, and denote the obfuscation scheme as MPr for midpoint protocol with rounding and MPSSr for midpoint protocol with sub-sampling and rounding.

We also consider two alternative methods for obfuscation. First, the *item average* (IA) scheme replaces a user's rating with the average rating of the item, computed from the training set. Second, the *feature average* (FA) scheme replaces the user's rating with the average rating provided by the feature classes (e.g., males and females), each with probability 0.5.

Finally, we evaluate each of the above obfuscation schemes, i.e., MP, MPr, IA and FA, together with sub-sampling (SS). As a baseline, we also evaluated the privacy risk and the prediction accuracy when *no obfuscation* scheme is used (NO).

7.2 Experimental Results

Dense Dataset. We begin by evaluating the obfuscation schemes on the dense PTV dataset using its two users' features (gender and political affiliation), illustrated in Figures 4a and 4b, respectively. Each figure shows the privacy risk (AUC) computed using the 4 inference methods, and the prediction accuracy (RMSE) on applying different obfuscation schemes.

Both figures clearly show that MP successfully mitigates the privacy risk (AUC is around 0.5) whereas the prediction accuracy is hardly impacted (2% increase in RMSE). This illustrates that MP attains excellent privacy in practice, and that our modeling assumptions are reasonable: there is little correlation to the private feature after the category bias is removed. Indeed, strong cor-

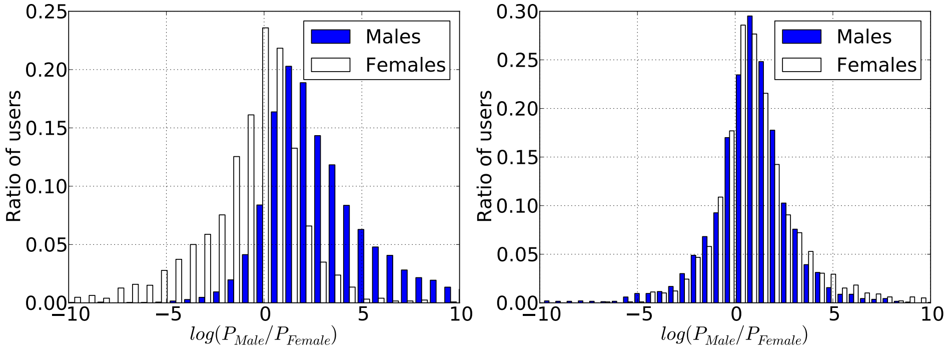


Figure 6: Distribution of inference probabilities for males and females using Movielens dataset and logistic regression (left) before obfuscation and (right) after MPSS obfuscation.

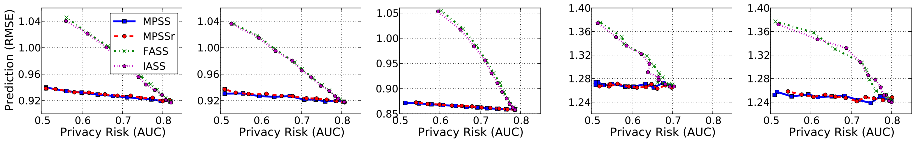
relations not captured by (7) could manifest as failure to block inference after obfuscation, especially through the non-linear SVM classifier. This is clearly not the case, indicating that any dependence on the private feature not captured by (7) is quite weak.

Adding rounding (MPr), which is essential for real-world deployment of MP, has very little effect on both the AUC and RMSE. Though IA and FA are successful in mitigating the privacy risk, they are suboptimal in terms of prediction. They severely impact the prediction accuracy, increasing the RMSE by roughly 9%. Finally, since this is a dense dataset, there is little correlation between the private feature and the set of items a user rates. Therefore, MP without SS suffices to mitigate the privacy risk.

Sparse Datasets. Next, we investigate the effect of partial feedback by evaluating our obfuscation schemes on the Movielens and Flixster datasets. In these datasets, in addition to the rating value, the set of items rated by a user can be correlated with her private feature. The results for obfuscation on Movielens and Flixster are in Figure 5.

For all three datasets, we observe that MP successfully blocks inference by LSE, but fails against the other three methods. This is expected, as the items rated are correlated to the private feature, and LSE is the only method among the four that is insensitive to this set. For the same reason, SS *alone* defeats all methods *except* LSE, which still detects the feature from the unobfuscated ratings (AUC 0.69–0.71). Finally, MPSS and MPSSr have excellent performance across the board, both in terms of privacy risk (AUC 0.5–0.55) and impact on prediction accuracy (up to 5%). In contrast, IA and FA significantly increase the RMSE (around 15%). We stress that, in these datasets, items are not rated independently as postulated by (18). Nevertheless, the results above indicate that MPSS blocks inference in practice *even when this assumption is relaxed.*

We further quantify the impact of sub-sampling in terms of the number of items that are not reported to the analyst in a partial feedback setting. To this end, we compute the ratio of items excluded in the feedback reported by



(a) Movielens Gender (b) Movielens Age (c) Flixster Gender (d) PTV Gender (e) PTV Politics

Figure 7: Prediction accuracy (RMSE) vs. privacy risk (LSE AUC) tradeoff for varying levels of obfuscation. Our proposed schemes (MPSS and MPSSr) have little impact on prediction accuracy as privacy is increased, whereas the prediction accuracy worsens dramatically under the baseline schemes.

the user as a result of applying SS. We found that for the dense PTV dataset, 80% of the users include all their ratings in their partial feedback, and the remaining 20% exclude at most 5% of their ratings. For the sparse Flixster and Movielens datasets, 50% of the users do not include 10% and 23% of their ratings, respectively. All users include at least 50% of their ratings, hence the prediction accuracy does not suffer with MPSS obfuscation.

Overall, these results indicate that both MP and MPSS are highly effective in real-world datasets – they mitigate the privacy risk while incurring very small impact on the prediction accuracy. Moreover, these obfuscation schemes work well even when facing non-linear inference methods, such as SVM, indicating that in practice, they eliminate any dependency between the ratings and the private feature.

Privacy Risk. To further illustrate how obfuscation defeats the inference of a private feature, we focus of the effect of obfuscation on logistic regression over the Movielens Gender dataset. Figures 6(a) and 6(b) plot the distribution of $\log(P_{Male}/P_{Female})$ (a) before obfuscation and (b) after obfuscation with MPSS. Here, P_{Male} and P_{Female} are the posterior probabilities for the two classes as obtained through logistic regression. Prior to obfuscation, there is a clear separation between the distributions of males and females, enabling successful gender inference (AUC 0.82 as shown in Figure 5a). However, after obfuscation, the two distributions become indistinguishable (AUC 0.54).

Privacy-Accuracy Tradeoff. Finally, we study the privacy-prediction accuracy tradeoff by applying an obfuscation scheme on an item rating with probability α, and releasing the real rating with probability $1-\alpha$. We vary the value of α between 0 and 1 in steps of 0.1, that is, when $\alpha = 0$ no obfuscation is performed, and $\alpha = 1$ means that all ratings are obfuscated. For each α, we measure the RMSE as well as the AUC of LSE.

Figure 7 shows the resulting RMSE-AUC tradeoff curves for MPSS, MPSSr and the two baseline obfuscation schemes with sub-sampling. The figure shows that MPSS and MPSSr provide the best privacy-accuracy tradeoff (the slopes of the curves are almost flat), and consistently obtain better prediction accuracy (lower RMSE) for the same privacy risk (inference AUC) than all other methods.

8 Conclusion

We have introduced a framework for reasoning about privacy, accuracy, and dislosure tradeoffs in matrix factorization. This naturally raises the question of how these tradeoffs extend to other statistical or prediction tasks. An orthogonal direction to the one we pursued, when seeking a mininal disclosure, is to investigate schemes that are not perfectly private. It would be interesting to investigate, e.g., privacy-dislosure tradeoffs, rather than the usual privacy-accuracy tradeoffs one encounters in literature. For example, it is not clear whether one can construct protocols in which the distribution of the obfuscated output differs accross users with opposite private attribute by, e.g., an ϵ factor, but leak less information than MP: such protocols could, e.g., disclose a quantized version of the biases for each item.

References

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. *ACM Sigmod Record*, 2000.

[2] S. Banerjee, N. Hegde, and L. Massoulié. The price of privacy in untrusted recommendation engines. In *Allerton*, 2012.

[3] S. Bhagat, I. Rozenbaum, and G. Cormode. Applying link-based classification to label blogs. In *WebKDD*, 2007.

[4] S. Bhagat, U. Weinsberg, S. Ioannidis, and N. Taft. Recommending with an agenda: Active learning of private attributes using matrix factorization, 2013. http://arxiv.org/abs/1311.6802.

[5] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users. In *WWW*, 2013.

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

[7] F. Calmon and N. Fawaz. Privacy against statistical inference. In *Allerton*, 2012.

[8] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 2010.

[9] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 2009.

[10] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *JMLR*, 2011.

[11] K. Chaudhuri and A. D. Sarwate. A near-optimal algorithm for differentially-private principal components. *JMLR*, 2013.

[12] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, 2013.

[13] C. Dwork. Differential privacy. In *ICALP*. 2006.

[14] C. Dwork and J. Lei. Differential privacy and robust statistics. In *STOC*, 2009.

[15] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[16] S. Friedland and S. Gaubert. Submodular spectral functions of principal submatrices of a hermitian matrix, extensions and applications. *Linear Algebra and its Applications*, 2013.

[17] S. Goldwasser and M. Bellare. *Lecture Notes on Cryptography.* MIT, 2001.

[18] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 743:29–36, 1982.

[19] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2nd edition, 2009.

[20] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *RecSys*, 2010.

[21] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 2011.

[22] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *JMLR*, 2010.

[23] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, 2008.

[24] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 2009.

[25] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 2013.

[26] E. L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer, 2nd edition, 1998.

[27] A. Machanavajjhala, A. Korolova, and A. D. Sarma. Personalized social recommendations - accurate or private? *PVLDB*, 2011.

[28] A. Makhdoumi and N. Fawaz. Privacy-utility tradeoff under statistical uncertainty. In *Allerton*, 2013.

[29] F. McSherry and I. Mironov. Differentially private recommender systems: building privacy into the net. In *KDD*, 2009.

[30] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in Online Social Networks. In *WSDM*, 2010.

[31] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *OAKLAND*, 2008.

[32] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions–i. *Mathematical Programming*, 14(1):265–294, 1978.

[33] S. R. Oliveira and O. R. Zaiane. Privacy preserving clustering by data transformation. In *SBBD*, pages 304–318, 2003.

[34] J. Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *CIKM*, 2010.

[35] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *SMUC*, 2010.

[36] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *IEEE TKDE*, 2010.

[37] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, , and N. Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *Journal of Privacy and Confidentiality*, 2012.

[38] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft. How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data. In *GlobalSIP*, 2013.

[39] L. Sankar, S. R. Rajagopalan, and H. V. Poor. Utility-privacy tradeoff in databases: An information-theoretic approach. *IEEE Trans. Information Forensics and Security*, 2013.

[40] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *STOC*, 2011.

[41] H. Strasser. Local asymptotic minimax properties of Pitman estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 60(2):223–247, 1982.

[42] J. Vaidya, C. W. Clifton, and Y. M. Zhu. *Privacy Preserving Data Mining.* Springer, 2006.

[43] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 1965.

[44] U. Weinsberg, S. Bhagat, S. Ioannidis, and N. Taft. Blurme: Inferring and obfuscating user gender based on ratings. In *RecSys*, 2012.

[45] H. Yamamoto. A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers. *IEEE Trans. Inform. Theory*, 1983.

A Optimality of MPSS

We begin by defining the class of learning protocols within which we will show that MPSS is optimal.

Learning Protocols. We define a learning protocol as a tuple $\mathcal{R} = (L, (\mathcal{S}_R, Y), \widehat{\mathbf{x}})$ where:

- The disclosure $L : \mathbb{R}^{d+1}_{-\mathbf{0}} \times [0,1] \times [0,1] \to \mathcal{L}$ at each item $j \in \mathcal{S}$ is now a function of the extended profiles *and* the rating probabilities, i.e., $\ell_j = L(v_j, p_j^+, p_j^-)$, $j \in \mathcal{S}$. We again denote by $\ell = L(\mathcal{V}, p) \in \mathcal{L}^{|S|}$ the vector of disclosures.
- The obfuscated user feedback is constructed in two steps. First, the user computes a set $\mathcal{S}_R(S_0, x_0, \ell) \subseteq S_0$, which determines the items for which the she will reveal her rating to the analyst; second, having determined $\mathcal{S}_R$, the user produces an obfuscated output $y = Y(r_{S_R}, x_0, \ell)$, where $r_{S_R} \in \mathbb{R}^{|S_R|}$ the vector of ratings for items in set $\mathcal{S}_R$. Note that $\mathcal{S}_R$ is constrained to be a subset of $\mathcal{S}_0$: the user may only reveal ratings for a subset of the items she has truly rated. The feedback of the user to the analyst is the pair $(\mathcal{S}_R, y)$, i.e., the user reveals along with the feedback $y \in \mathcal{Y}$ the items for which she provides feedback. Formally, these two are determined by a mapping $\mathcal{S}_R : 2^{\mathcal{S}} \times \{-1, +1\} \times \mathcal{L}^{|\mathcal{S}|} \to 2^{\mathcal{S}}$, that determines the set $\mathcal{S}_R \subseteq \mathcal{S}_0$, and a family of mappings $Y : \mathbb{R}^{|\mathcal{S}_R|} \times \{-1, +1\} \times \mathcal{L}^{|\mathcal{S}_R|} \to \mathcal{Y}$, one for each $\mathcal{S}_R \subseteq \mathcal{S}$.

- The estimator $\widehat{\mathbf{x}} = \mathbf{x}((\mathcal{S}_R, y), \mathcal{V})$ is now a mapping $\widehat{\mathbf{x}} : 2^{\mathcal{S}} \times \mathcal{Y} \times (\mathbb{R}^{d+1}_{-\mathbf{0}})^{|\mathcal{S}|}$ that depends on the user's feedback $(\mathcal{S}_R, y)$, as well as the profile information available to the analyst.

We can naturally define partial orderings of learning protocols $\mathcal{R} = (L, (\mathcal{S}_R, Y), \widehat{\mathbf{x}})$ with respect to the extent of disclosure by extending Definition 3 in a straightforward fashion. Regarding accuracy, we say that $\mathcal{R}$ is more accurate than $\mathcal{R}'$ if it yields a smaller expected ℓ_2 loss *conditioned on* $\mathcal{S}_0$. Finally, regarding privacy, we say that $\mathcal{R}$ is privacy-preserving if the joint distribution of the random variables $(\mathcal{S}_R, y)$ does not depend on x_0: both the set $\mathcal{R}$, as well as the corresponding obfuscated feedback y, are equal in distribution when $x_0 = +1$ or $x_0 = -1$.

We will further restrict our analysis to protocols that satisfy the following property.

Definition 4. *Let*

$$\mathcal{S}^+ = \{j \in \mathcal{S} : \rho_j \leq 1\}, \quad \mathcal{S}^- = \{j \in \mathcal{S} : \rho_j > 1\}, \tag{24}$$

be the set of items more likely to be rated by "positive" and "negative" users respectively. We say that $\mathcal{R} = (L, (\mathcal{S}_R, Y), \widehat{\mathbf{x}})$ *is* positive-negative independent *(PNI) if the random sets* $\mathcal{S}_R \cap \mathcal{S}_+$ *and* $\mathcal{S}_R \cap \mathcal{S}_-$ *are independent random variables.*

Note that MPSS is a PNI protocol, and so is any protocol in which the events $\{j \in \mathcal{S}_R\}$ are independent Bernoulli variables for every $j \in \mathcal{S}$.

Optimality. The following theorem holds

Theorem 2. *Under* (7) *with Gaussian noise, and* (18)*:*

1. MPSS is privacy preserving.

2. There is no privacy preserving, PNI learning protocol that is strictly more accurate than MPSS.

3. Any privacy preserving, PNI learning protocol that does not disclose as much information as MPSS must also be strictly less accurate.

Proof. We begin our proof of Theorem 2 by establishing a few auxiliary results. Denote by $\mathcal{R} = (L, (\mathcal{S}_R, Y), \widehat{\mathbf{x}})$ be the MPSS protocol. Our first lemma states that (20) is an upper bound among privacy preserving protocols:

Lemma 2. *Let* $\mathcal{R}' = (L', (\mathcal{S}'_R, Y'), \widehat{\mathbf{x}}')$ *be privacy-preserving. Then* $\mathrm{P}_{x,\mathcal{V},p}(j \in \mathcal{S}'_R) \leq \min(p^+_j, p^-_j)$.

Proof. Recall that, by construction $\mathcal{S}'_R \subseteq \mathcal{S}_0$, the actual items rated by a user. Hence $\mathrm{P}_{(+1,\mathbf{x}),cV,p}(j \in \mathcal{S}'_R) \leq p^+_j$ and $\mathrm{P}_{(-1,\mathbf{x}),cV,p}(j \in \mathcal{S}'_R) \leq p^-_j$. As $\mathcal{R}$ is privacy preserving, these l.h.s. probabilities are equal, and the lemma follows. □

In fact, this inequality becomes strict if L' does not disclose $\rho_j = p^-_j / p^+_j$, for some $j \in \mathcal{S}$.

Lemma 3. *Let $\mathcal{R}' = (L', (\mathcal{S}'_R, Y'), \widehat{\mathbf{x}}')$ be privacy-preserving, and suppose that L' does not disclose $\rho = p^-/p^+$–i.e., there is no $\phi : \mathcal{L}' \to \mathbb{R}$ such that $p^-/p^+ = \phi(L'(v, p^+, p^-))$ for all v, p^+, p^-. Then, there exist values $p^+, p^- \in [0,1]$, an extended profile $v \in \mathbb{R}^{d+1}_{-\mathbf{0}}$, and an $x_0 \in \{-1, +1\}$ such that $\mathrm{P}_{x,\mathcal{V},p}(j \in \mathcal{S}_R) < \min(p^+_j, p^+_j)$ for all $\mathcal{V}$ and p such that $v_j = v$ and $(p^+_j, p^-_j) = (p^+, p^-)$.*

Proof. Assumption that $\mathcal{R}'$ does not disclose ρ_j, for some $j \in \mathcal{S}$. Then, there exist probabilities $p^+, q^+, p^-, q^- \in [0,1]$ and extended vectors $v, v' \in \mathbb{R}^{d+1}_{-\mathbf{0}}$ such that

$$\rho \equiv p^-/p^+ < q^-/q^+ \equiv \rho',$$

while $L(v, p^+, p^-) = L(v', q^+, q^-)$. Consider any two $\mathcal{V}, \mathcal{V}' \subseteq \mathbb{R}^{d+1}_{-\mathbf{0}}$ and $p, p' \in ([0,1] \times [0,1])^{|\mathcal{S}|}$ such that all item profiles and probabilities are identical for all $k\mathcal{S}$, but differ in j: the j-th elements in $\mathcal{V}, p$ are v and (p^+, p^-), respectively, while the j-th elements of $\mathcal{V}', p'$ are v' and (q^+, q^-), respectively. Observe that $\equiv L'(\mathcal{V}, p) = L(\mathcal{V}', p')$.

Recall that $\mathcal{S}'_R$ depends on $\mathcal{S}_0$, x_0, and the disclosure from the analyst. Hence, as $\equiv L'(\mathcal{V}, p) = L(\mathcal{V}', p')$, conditioned on $\mathcal{S}_0$, $\mathcal{S}'_R$ is identically distributed in both cases. In particular,

$$\mathrm{P}_{x,\mathcal{V},p}(j \in \mathcal{S}'_R \mid \mathcal{S}_0 = \mathcal{A}) = \mathrm{P}_{x,\mathcal{V}',p'}(j \in \mathcal{S}'_R \mid \mathcal{S}_\prime = \mathcal{A}), \tag{25}$$

for all $\mathcal{A} \subseteq \mathcal{S}$. As $\mathcal{S}'_R \subseteq \mathcal{S}_0$, we have $\mathrm{P}_{x,\mathcal{V},p}(j = \mathcal{S}'_R) \overset{(18)}{=} Z \cdot p_j^{x_0}$ where

$$Z_{x,\mathcal{V},p} = \sum_{\mathcal{A} \subseteq \mathcal{S} \setminus \{j\}} \mathrm{P}_{x,\mathcal{V},p}(j \in \mathcal{S}'_R | \mathcal{S}_0 = \mathcal{A} \cup \{j\}) \prod_{k \in \mathcal{A}} p_k^{x_0} \prod_{k \in \mathcal{S} \setminus (\mathcal{A} \cup \{j\})} (1 - p_k^{x_0})$$

As $\mathcal{R}'$ is privacy preserving, by Lemma 2 we get that $Z \leq \min(1, \rho^{x_0})$. Repeating the same steps for $\mathrm{P}_{x,\mathcal{V}',p'}(j = \mathcal{S}'_R)$, we get that also $Z_{x,\mathcal{V}',p'} \leq \min(1, (\rho')^{x_0}\}$. By (25), these are equal, and thus

$$Z = Z_{x,\mathcal{V},p} = Z_{x,\mathcal{V}',p'} \leq \min(1, (\rho)^{x_0}, (\rho')^{x_0})$$

Recall that $\rho < \rho'$, by construction. If $\rho < 1$, then for $x_0 = +1$ we get $\min(1, (\rho)^{x_0}, (\rho')^{x_0}) = \rho$. Then,

$$\mathrm{P}_{(+1,\mathbf{x}),\mathcal{V}',p'}(j = \mathcal{S}'_R) = Zq^+ \leq \rho q^+ < \min(1, \rho')q^+ = \min(q^+, q^-)$$

and the lemma holds for $x_0 = +1$, v', and (q^+, q^-). If $\rho \geq 1$, then for $x_0 = -1$ we get $\min(1, (\rho)^{x_0}, (\rho')^{x_0}) = (\rho')^{-1}$, and

$$\mathrm{P}_{(-1,\mathbf{x}),\mathcal{V},p}(j = \mathcal{S}'_R) = Zp^- \leq p^-/\rho' < \min(1, \rho^{-1})p^- = \min(p^+, p^-)$$

and the lemma holds for $x_0 = -1$, v, and (p^+, p^-). □

The PNI property allows us to couple $\mathcal{S}'_R$ and $\mathcal{S}_R$ in a way that the latter dominates the former.

Lemma 4. *Let* $\mathcal{R} = (L, (\mathcal{S}_R, Y), \widehat{\mathbf{x}})$ *be the MPSS protocol, and* $\mathcal{R}' = (L', (\mathcal{S}'_R, Y'), \widehat{\mathbf{x}}')$ *a privacy-preserving, PNI protocol. Then, there exists a joint probability space in which* $\mathcal{S}'_R \subseteq \mathcal{S}_R \subset \mathcal{S}_0$.

Proof. Recall that $\mathcal{S} = \mathcal{S}^+ \cup \mathcal{S}^-$, where $\mathcal{S}^+, \mathcal{S}^-$ the sets of positive and negative items in (24). Let $\mathcal{S}_{0+}$ and $\mathcal{S}_{0-}$ the set of items rated by two users of type $x_0 = +1$ and $x_0 = -1$, respectively. We construct $\mathcal{S}_{0+}, \mathcal{S}_{0-}$ on the same probability space as follows: for each j, draw X_j uniform in [0,1], and let $j \in \mathcal{S}_{0+}$ iff $X_j \leq p_j^+$ and $j \in \mathcal{S}_{0-}$ iff $X_i \leq p_i^-$. The sets $\mathcal{S}_{0+}, \mathcal{S}_{0-}$ can be intersected in the obvious way with M(+) and M(-) yielding

$$\mathcal{S}_{0+} = \mathcal{S}_{0+}^+ \cup \mathcal{S}_{0+}^-, \quad \mathcal{S}_{0-} = \mathcal{S}_{0-}^+ \cup \mathcal{S}_{0-}^-$$

Then we have, a.s. $\mathcal{S}_{0+}^+ \supseteq \mathcal{S}_{0-}^+$ and $\mathcal{S}_{0+}^- \subseteq \mathcal{S}_{0-}^-$ Now, we can construct the set $\mathcal{S}_R$ reported by MPSS on the same space by letting $\mathcal{S}_R = \mathcal{S}_R^+ \cup \mathcal{S}_R^-$ where $\mathcal{S}_R^+ = \mathcal{S}_{0-}^+$ and $\mathcal{S}_R^- = \mathcal{S}_{0+}^-$.

Now apply any privacy preserving mechanism $\mathcal{R}'$ to $\mathcal{S}_{0+}$ and $\mathcal{S}_{0-}$. This will yield sets $\mathcal{Q}_{R+}$, $\mathcal{Q}_{R-}$, that can also be decomposed as above:

$$\mathcal{Q}_{R+} = \mathcal{Q}_{R+}^+ \cup \mathcal{Q}_{R+}^-, \quad \mathcal{Q}_{R-} = \mathcal{Q}_{R-}^+ \cup \mathcal{Q}_{R-}^-$$

The sets $\mathcal{Q}_{R+}$, $\mathcal{Q}_{R-}$, are not necessarily equal, but must satisfy the following the properties:

$$\mathcal{Q}_{R-}^+ \subseteq \mathcal{S}_{0-}^+ = \mathcal{S}_R^+, \ \mathcal{Q}_{R+}^- \subseteq \mathcal{S}_{0+}^- = \mathcal{S}_R^-, \text{ (by construction)}, \tag{26}$$

$$\mathcal{Q}_{R-}^+ \overset{\mathrm{d}}{=} \mathcal{Q}_{R+}^+, \quad \mathcal{Q}_{R-}^- \overset{\mathrm{d}}{=} \mathcal{Q}_{R+}^-, \text{ (by privacy)} \tag{27}$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution. Define $\mathcal{S}'_R \equiv \mathcal{Q}_{R-}^+ \cup \mathcal{Q}_{R+}^-$. By (26), we get $\mathcal{S}'_R \subseteq \mathcal{S}_R$ with probability 1. Moreover, by (27) and the fact that $\mathcal{R}'$ is PNI, we get that $\mathcal{S}'_R \overset{\mathrm{d}}{=} \mathcal{Q}_{R+} \overset{\mathrm{d}}{=} \mathcal{Q}_{R-}$. □

We are ready to prove Theorem 2. Privacy follows directly from (20). Theorem 1 implies MPSS yields minimal ℓ_2 loss conditioned on $\mathcal{S}_R$. Optimality conditioned on $\mathcal{S}_0$ follows from Lemma 4, and the fact that the ℓ_2 loss (12) is a monotone decreasing function of $\mathcal{S}_R$. Finally, any protocol that does not does not disclose v_{j0}, for some $j \in \mathcal{S}$ will lead to a higher loss by Theorem 1. Moreover, by Lemmas 3 and 4, if a protocol $\mathcal{R}'$ does not disclose $\rho_j = p_j^-/p_j^+$ for some $j \in \mathcal{S}$, there exist x_0, $\mathcal{V}$, and p for which one can construct a coupling of $\mathcal{S}'_R$ and $\mathcal{S}_R$ such that $\mathcal{S}'_R \subset \mathcal{S}_R$ with non zero probability; the minimality of the disclosure therefore follows, again from the fact that the ℓ_2 loss (12) is decreasing in $\mathcal{S}_R$. □