BlurMe: Inferring and Obfuscating User Gender Based on Ratings

Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, Nina Taft
{udi.weinsberg, smriti.bhagat, stratis.ioannidis, nina.taft}@technicolor.com
Technicolor
Palo Alto

ABSTRACT

User demographics, such as age, gender and ethnicity, are routinely used for targeting content and advertising products to users. Similarly, recommender systems utilize user demographics for personalizing recommendations and overcoming the cold-start problem. Often, privacy-concerned users do not provide these details in their online profiles. In this work, we show that a recommender system can infer the gender of a user with high accuracy, based solely on the ratings provided by users (without additional metadata), and a relatively small number of users who share their demographics. We design techniques for effectively adding ratings to a user's profile for obfuscating the user's gender, while having an insignificant effect on the recommendations provided to that user.

Categories and Subject Descriptors: H.2.8 Database Applications: Data Mining
Keywords: Recommender Systems, Privacy.

1. INTRODUCTION

Profiling users through demographic information, such as gender, age, or ethnicity, is of great importance in targeted advertising and personalized content delivery. Recommender systems too can benefit from such information to provide personalized recommendations. However, users of recommender systems often do not volunteer this information. This may be intentional – to protect their privacy, or unintentional – out of laziness or disinterest. As such, traditional collaborative filtering methods eschew using such information, relying instead solely on ratings provided by users.

At a first glance, disclosing ratings to a recommender system may appear as a rather innocuous action. There is certainly a utility users accrue from this disclosure – namely, the ability to discover relevant items. Nevertheless, there has been a fair amount of work indicating that user demographics are correlated to, and thus can be inferred from, user activity on social networks [9], blogs [2], and microblogs [12] etc. It is thus natural to ask whether demographic information such as age, gender, ethnicity or even political orientation can also be inferred from information disclosed to recommender systems. Indeed, irrespective of a rating value, the mere fact that a user has interacted with an item (*e.g.*, viewed a specific movie or purchased a product) may be correlated with demographic information.

The potential success of such an inference has several important implications. From the recommender's perspective, profiling users not only improves their own recommendations, but also enables targeted advertising. From the user's perspective, the success of such an inference raises serious privacy concerns. A privacy-conscious user cannot simply withhold all information as this would come at the cost of foregoing the utility gained from using the recommender system in the first place – namely, finding relevant content. Explicitly withholding the user's demographic information does not ensure privacy either, as it may be possible to uncover it through inference. Because the approach of withholding information is often impractical, we believe a more promising approach is that of adding ratings into a user profile with the intent of creating ambiguity. In this paper, we thus explore both the questions of how demographic information can be inferred – from ratings data alone – and that of how to hinder such inference via obfuscating the information disclosed to the recommender system.

In general, any *obfuscation mechanism* employed by the user strikes a tradeoff. This is between (a) the user's privacy, as captured by the recommender's ability to infer her demographic information, and (b) the utility to the user, captured by the accuracy of recommendations she receives. Understanding the nature of such a tradeoff is thus a fundamental question. In this work, we study the above issues in a comprehensive manner, making the following contributions:

- We evaluate several gender inference algorithms on two movie ratings datasets, Movielens and Flixster, and show that a relatively small amount of labeled data (*i.e.*, users who share their gender), is sufficient to predict the gender of users with about 80% accuracy.
- We find that the act of watching a movie, regardless of the rating given, is strongly correlated with one's gender, and we identify movies for which this correlation is high.
- Based on these observations, we propose several obfuscation mechanisms, allowing the users to alter the information they reveal to the recommender service.
- We further evaluate these mechanisms with respect to the trade-offs they achieve between user privacy, as captured by the accuracy of gender-inference, versus user utility, as captured by rating prediction RMSE.

RecSys'12 September 9-13, 2012, Dublin, Ireland

- We establish that quite favorable tradeoffs are feasible; indeed, using 1% additional ratings it is possible to reduce the success of gender inference by 80% while reducing the quality of recommendations only by 1.3%.

To the best of our knowledge, we are the first to study and quantify demographic inference methods that rely solely on rating data. Moreover, we are the first to design and analyze obfuscation mechanisms aiming to preserve gender privacy while maintaining recommendation accuracy.

2. RELATED WORK

Inferring demographics of users has been widely studied in different contexts, and for various types of user-generated data. In the context of interaction networks, the graph structure has been shown to be useful for inferring demographics using link-based information for blog [2] and social network [9] data from Facebook. Other works rely on the textual features derived from writings of users to infer demographics. For instance, Rao et al. [12] use an SVM classifier on Twitter data, and Otterbacher et al. use logistic regression on movie reviews from IMDB [11]. It is useful to note that the prediction accuracy obtained using logistic regression on movie reviews is about 73.7%, lower than that obtained using the same algorithm on movie ratings, albeit for a different dataset. In our setting, the input to the gender inference mechanism is only the movie ratings provided by users, with no metadata about movies or users. While there has been work on collective matrix factorization [13] to take into account attributes of movies and users in addition to ratings for making recommendations, the work does not explore the specific task of inferring user demographics.

Rather than focusing solely on inference, our goal is to be able to use insights gained to design mechanisms that obfuscate users' demographics. Injecting noise for privacy was recently studied in [15, 14] for search privacy, where the goal is to obfuscate search engine queries rather than a user's demographics.

Our work is also related to studies of robustness in recommender systems [10, 3, 1]. The goal of such studies is to evaluate how an attacker can manipulate a recommender system by injecting adversarially selected ratings. In contrast, our study is based on the interaction of a user, whose ratings are not necessarily added in the training set, with the recommender system. Although the user may submit altered ratings, her interest is still in receiving relevant recommendations, albeit without disclosing her gender.

An elegant and formal approach to privacy in recommender systems has been made through differential privacy [8]. Nevertheless, differential privacy guarantees aim at a different goal, which is to ensure that the output of a recommender depends only marginally on the input of any single user. In contrast, we aim at not protecting ratings per se, but the demographic information of each user; this notion cannot be captured within the formalism of differential privacy.

3. PROBLEM DEFINITION

For the sake of concreteness, we assume throughout the paper that the information users wish to protect is their gender; nevertheless, our algorithms are generic, and apply also when different demographic features (age, ethnicity, political orientation, *etc.*), expressed as a categorical variable, are to be protected.

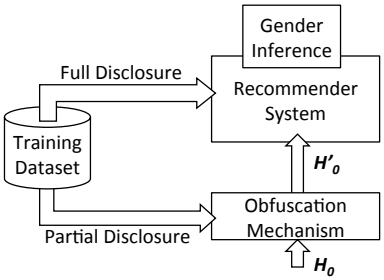


Figure 1: Illustration of the problem setup

3.1 Setup

Our setup is summarized in Figure 1. A user, indexed by 0, views and rates items which, for concreteness, we refer to as movies. We assume that the universe of movies the user can rate comprises a catalog of M movies; the user rates a subset $\mathcal{S}_0$ of the catalog $\mathcal{M} = \{1, 2, \ldots, M\}$. We denote by $r_{0j} \in \mathbb{R}$ the rating of movie $j \in \mathcal{S}_0$ and define the user's *rating profile* as the set of (movie, rating) pairs $\mathcal{H}_0 \equiv \{(j, r_{0j}) : j \in \mathcal{S}_0\}$. The user submits $\mathcal{H}_0$ to an *obfuscation mechanism*, which outputs an altered rating profile $\mathcal{H}'_0 = \{(j, r'_{0j}) : j \in \mathcal{S}'_0\}$, for some $\mathcal{S}'_0 \neq \mathcal{S}_0$. In simple terms, this obfuscation aims at striking a good balance between the following two conflicting goals : (a) $\mathcal{H}'_0$ can be used to provide relevant recommendations to the user, and (b) it is difficult to infer the user's gender from $\mathcal{H}'_0$.

More specifically, we assume that the obfuscated rating profile $\mathcal{H}'_0$ is submitted to a *recommender* mechanism that has a module that implements a *gender inference* mechanism. The recommender mechanism uses $\mathcal{H}'_0$ to *predict* the user's ratings on $\mathcal{M} \setminus \mathcal{S}'_0$, and potentially, recommend movies that might be of interest to the user. The gender inference module is a classification mechanism, that uses the same $\mathcal{H}'_0$ to profile and label the user as either male of female.

Though the implementation of the recommender mechanism might be publicly known, the obfuscation and gender inference mechanisms are not. As a first step in this problem, we take the simple approach that both recommendation and gender inference are oblivious to the fact that any kind of obfuscation is taking place. Both mechanisms take the profile $\mathcal{H}'$ at "face value" and do not reverse-engineer the "true" profile $\mathcal{H}$. (We leave for future work the case when the obfuscation mechanism is known.)

3.2 Training Dataset

We assume that the recommender and inference mechanisms have access to a *training dataset*. This dataset comprises a set of $\mathcal{N} = \{1, \ldots, N\}$ users each of which has given ratings to a subset of the movies in the catalog $\mathcal{M}$. We denote by $\mathcal{S}_i \subseteq \mathcal{M}$ the set of movies for which the rating of a user $i \in \mathcal{N}$ is in the dataset, and by r_{ij}, $j \in \mathcal{S}_i$, the rating given by user $i \in \mathcal{N}$ to movie $j \in \mathcal{M}$. Moreover, for each $i \in \mathcal{N}$ the training set also contains a binary variable $y_i \in \{0, 1\}$ indicating the gender of the user (we map bit 0 to male users). We assume that the training set is unadulterated: neither ratings nor gender labels have been tampered with or obfuscated.

The obfuscation mechanism may also have a partial view of the training set. In the extreme case, the training dataset is public, and the obfuscation mechanism has full access to it. It is interesting however to consider weaker obfuscation

mechanisms, that can only access limited statistics (or other queries over the dataset), such as, the average rating of a movie. Though the mechanisms we propose can, *a fortiori*, be implemented when the dataset is public, we will state the training set statistics required in their implementation.

3.3 Matrix Factorization

The main focus of this paper is the design and analysis of mechanisms for gender inference and obfuscation. As such, we fix the recommender mechanism throughout the paper to be matrix factorization [6], since this is commonly used in commercial systems. In short, given the rating profile $\mathcal{H}'_0$, we generate ratings for the set $\mathcal{M} \setminus \mathcal{S}_0$ by appending the provided ratings to the rating matrix of the training set and factorizing it.

More specifically, we associate with each user $i \in \mathcal{N} \cup \{0\}$ a latent feature vector $u_i \in \mathbb{R}^d$. We also associate with each movie $j \in \mathcal{M}$ a latent feature vector $v_j \in \mathbb{R}^d$. We define the regularized mean square error to be

$$\sum_{i \in \mathcal{N} \cup \{0\}, j \in \mathcal{S}_i} (r_{i,j} - \langle u_i, v_j \rangle - \mu)^2 + \lambda \sum_{i \in \mathcal{N} \cup \{0\}} \|u_i\|_2^2 + \lambda \sum_{j \in \mathcal{M}} \|v_j\|_2^2$$

where μ is the average rating of the entire dataset. We construct the vectors u_i, v_j by minimizing the MSE through gradient descent. We use $d = 20$ and $\lambda = 0.3$. Having profiled thusly both users and movies, we predict the rating of user 0 for movie $j \in \mathcal{M} \setminus \mathcal{S}'_0$ through $\langle u_0, v_j \rangle + \mu$.

3.4 Data Description

Flixster. Flixster is an online social network for rating and reviewing movies. Flixster allows users to enter demographic information into their profiles and share their movie ratings and reviews with their friends and the public. The dataset collected by Jamali et al. [5] has 1M users, of which only 34.2K users share their age and gender. We evaluate our techniques on this subset of 34.2K users, who have rated 17K movies and provided 5.8M ratings. The 12.8K males and 21.4K females have provided 2.4M and 3.4M ratings, respectively. Flixster allows users to provide half star ratings, however, to be consistent across the evaluation datasets, we round up the ratings to be integers from 1 to 5.

Movielens. Our second dataset is Movielens from the Grouplens[1] research team. The dataset consists of 3.7K movies and 1M ratings by 6K users. The 4331 males and 1709 females provided 750K and 250K ratings, respectively.

4. GENDER INFERENCE

In this section, we investigate whether inferring a user's gender based on her ratings is indeed possible. We study several different classifiers and evaluate them using the Flixster and Movielens datasets. We use the results of this analysis to inform our design of obfuscation mechanisms (Section 5).

4.1 Classifiers

To train our classifiers, we associate with each user $i \in \mathcal{N}$ in the training set a characteristic vector $x_i \in \mathbb{R}^M$ such that $x_{ij} = r_{ij}$, if $j \in \mathcal{S}_i$ and $x_{ij} = 0$, otherwise. Recall that the binary variable y_i indicates user i's gender, which serves as the dependent variable of our classification. We denote by $X \in \mathbb{R}^{N \times M}$ the matrix of characteristic vectors, and by $Y \in \{0,1\}^N$ the vector of genders.

We use three different types of classifiers: Bayesian classifiers, support vector machines (SVM) and logistic regression. In the Bayesian setting, we studied several different generative models; for all models, we assume that points (x_i, y_i) are sampled independently from the same joint distribution $P(x, y)$. Given P, the predicted label $\hat{y} \in \{0, 1\}$ attributed to characteristic vector x is the one with maximum likelihood, *i.e.*,

$$\hat{y} = \arg\max_{y \in \{0,1\}} P(y|x) = \arg\max_{y \in \{0,1\}} P(x, y) \quad (1)$$

Class Priors. The class prior classification serves as a base-line method for assessing the performance of the other classifiers. Given a dataset with unevenly distributed gender classes of the population, this basic classification strategy is to classify all users as having the dominant gender. This is equivalent to using (1) under the generative model $P(y|x) = P(y)$, estimated from the training set as:

$$P(y) = |\{i \in \mathcal{N} : y_i = y\}|/N. \quad (2)$$

Bernoulli Naïve Bayes. Bernoulli Naïve Bayes is a simple method that ignores the actual rating value. In particular, it assumes that a user rates movies independently and the decision to rate or not is a Bernoulli random variable. Formally, given a characteristic vector x, we define the rating indicator vector $\tilde{x} \in \mathbb{R}^M$ to be such that $\tilde{x}_j = \mathbb{1}_{x_j > 0}$. This captures the movies for which a rating is provided. Assuming that $\tilde{x}_j$, $j \in \mathcal{M}$, are independent Bernoulli, the generative model is given by $P(x, y) = P(y) \prod_{j \in \mathcal{M}} P(\tilde{x}_j|y)$ where $P(y)$ is the class prior, as in (2), and the conditional $P(\tilde{x}_j|y)$ is computed from the training set as follows:

$$P(\tilde{x}_j|y) = |\{i \in \mathcal{N} : \tilde{x}_{ij} = \tilde{x}_j \wedge y_i = y\}|/|\{i : y_i = y\}| \quad (3)$$

Multinomial Naïve Bayes. A drawback of Bernoulli Naïve Bayes is that it ignores rating values. One way of incorporating them is through Multinomial Naïve Bayes, which is often applied to document classification tasks [7]. Intuitively, this method extends Bernoulli to positive integer values by treating, *e.g.* a five-star rating as 5 independent occurrences of the Bernoulli random variable. Movies that receive high ratings have thus a larger impact on the classification. Formally, the generative model is given by $P(x, y) = P(y) \prod_{j \in \mathcal{M}} P(x_j|y)$ where $P(x_j|y) = P(\tilde{x}_j|y)^{x_j}$, and $P(\tilde{x}_j|y)$ is computed from the training set through (3).

Mixed Naïve Bayes. We propose an alternative to Multinomial, which we refer to as Mixed Naïve Bayes. This model is based on the assumption that, users give normally distributed ratings. More specifically,

$$P(x_j|\tilde{x}_j = 1, y) = (2\pi\sigma_y^2)^{-1/2} e^{-(x_j - \mu_{yj})^2/2\sigma_y^2}. \quad (4)$$

For each movie j, we estimate the mean μ_{yj} from the dataset as the average rating of movie j given by users of gender y, and the variance σ_y^2 as the variance of all ratings given by users of gender y. The joint likelihood used in (1) is then given by $P(x, y) = P(y) \prod_{j \in \mathcal{M}} P(\tilde{x}_j|y) P(x_j|\tilde{x_j}, y)$ where $P(y)$, $P(\tilde{x}_j|y)$ are estimated through (2) and (3), respectively. The conditional $P(x_j|\tilde{x_j}, y)$ is given by (4) when a rating is provided (*i.e.*, $\tilde{x}_j = 1$) and, trivially, by $P(x_j = 0|\tilde{x_j} = 0, y) = 1$, when it is not.

[1] www.grouplens.com/node/73

	Flixster		Movielens	
	AUC	P/R	AUC	P/R
Class Prior	0.50	0.39/0.62	0.50	0.51/0.72
Bernoulli	0.72	0.70/0.70	0.81	0.79/0.76
Multinomial	0.75	0.71/0.71	0.84	0.80/0.76
Mixed	0.74	0.71/0.71	0.82	0.79/0.77
SVM	0.82	0.73/0.70	**0.86**	0.78/0.77
SVM ($\tilde{X}$)	0.80	0.72/0.70	0.85	0.78/0.77
Logistic	**0.84**	0.76/0.77	0.85	0.80/0.80
Logistic ($\tilde{X}$)	0.83	0.75/0.76	0.84	0.78/0.79

Table 1: Mean AUC, precision (P) and recall (R)

	Flixster		Movielens	
	Female	Male	Female	Male
Class Prior	0.62/1	0/0	0/0	0.72/1
Bernoulli	0.75/0.80	0.62/0.54	0.57/0.73	0.88/0.78
Multinomial	0.76/0.78	0.63/0.60	0.57/0.73	0.89/0.77
Mixed	0.76/0.81	0.64/0.57	0.57/0.74	0.88/0.78
SVM	0.70/0.95	0.77/0.30	0.80/0.28	0.78/0.97
SVM ($\tilde{X}$)	0.69/0.96	0.77/0.27	0.80/0.28	0.77/0.97
Logistic	0.79/0.85	0.71/0.62	0.69/0.56	0.84/0.90
Logistic ($\tilde{X}$)	0.77/0.87	0.72/0.57	0.73/0.40	0.80/0.94

Table 2: Per-gender precision and recall.

Logistic Regression. A significant drawback of all of the above Bayesian methods is that they assume that movie ratings are independent. To address that, we applied logistic regression. Recall that linear regression yields a set of coefficients $\beta = \{\beta_0, \beta_1, ..., \beta_M\}$. The classification of a user $i \in N$ with characteristic vector x_i is performed by first calculating the probability $p_i = (1+e^{-(\beta_0+\beta_1 x_{i1}+...+\beta_M x_{iM})})^{-1}$. The user is classified as a female if $p_i < 0.5$ and as a male otherwise. The value p_i also serves a confidence value for the classification of user i. One of great benefits of using logistic regression is that the coefficients β capture the extent of the correlation between each movie and the class. In our case, the large positive β_j indicates that movie j is correlated with class male, whereas small negative β_j indicates that movie j is correlated with class female. We select the regularization parameter so that we have at least 1000 movies correlated with each gender that have a non-zero coefficient.

SVM. Intuitively, SVM finds a hyperplane that separates users belonging to different genders in a way that minimizes the distance of incorrectly classified users from the hyperplane (for a thorough explanation on SVMs see [4]). SVM holds many of the advantages of logistic regression – it does not assume independence in the feature space and produces coefficients. Since our feature space (number of movies) is already quite large, we use linear SVMs in our evaluations. We performed a logarithmic search over the parameter space (C) and found that $C = 1$ gave the best results.

4.2 Evaluation

We evaluate all algorithms on both the Flixster and Movielens datasets. We use 10-fold cross validation and compute the average precision and recall for the two genders across all folds. Additionally, we compute the Area Under the Curve (AUC) using the mean Receiver Operating Characteristic (ROC) curve computed across the folds. For the ROC, the true positive ratio is computed as the ratio of males correctly classified out of the males in the dataset, and the false positive ratio is computed as the ratio incorrectly classified males out of the females in the dataset. The ROC curves are given in Figure 2a and Figure 2b. Table 1 provides a summary of the classification results for 3 metrics: AUC, precision and recall. Table 2 shows the same results separated per-gender.

We see from the ROC curves that SVM and logistic regression perform better, across both datasets, than any of the Bayesian models since the regression curves for SVM and logistic dominate the others. In particular, logistic regression performed the best for Flixster while SVM performed best for Movielens. The performance of the Bernoulli, mixed and multinomial models do not different significantly from one another. These findings are further confirmed via the AUC values in Table 1. This table also shows the weakness of the simple class prior model that is easily outperformed by all other methods.

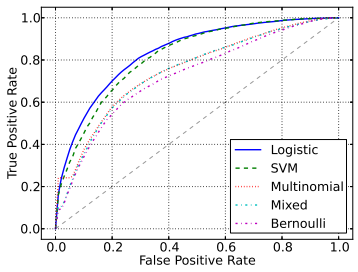
In terms of precision and recall, Table 2 shows that logistic regression outperforms all other models for Flixster users and both genders. For the Movielens users, SVM performs better than all other algorithms, while logistic regression is second best. In general, the inference performs better for the gender that is dominant in each dataset (female in Flixster and male in Movielens). This is especially evident for SVM, which exhibits very high recall for the dominate class and low recall for the dominated class. The mixed model improves significantly on the Bernoulli model and results similarly to the multinomial. This indicates that the usage of a Gaussian distribution might not be a sufficiently accurate estimation for the distribution of the ratings.

Impact of user ratings. We assess the importance of the rating value itself (number of stars) versus the simple binary event "watched or not" by applying logistic regression and SVM on a binary matrix, denoted by X, in which ratings are replaced by 1. Table 1 shows the performance of these two methods on X and $\tilde{X}$. Interestingly, SVM and logistic regression performed only slightly better when using X rather than $\tilde{X}$ as input, with less than 2% improvement on all measures. In fact, Table 2 indicates that although using X performs better than using $\tilde{X}$ for the dominant class, it is worse for the dominated class. Similarly, the Bernoulli model, which also ignores the rating values, performed relatively close to Multinomial and Mixed. This implies that whether or not a movie is included in one's profile is nearly as impactful as the value of star rating given for the movie. This has important ramifications for obfuscation mechanisms that need to do two things: decide which movies to add to a user profile, and decide which rating to give a movie. This finding suggests that the choice of which movies to add could have a large impact on impeding gender inference. However if the actual ratings do not impact gender inference much, then we could select a rating value that helps maintain the quality of recommendations.

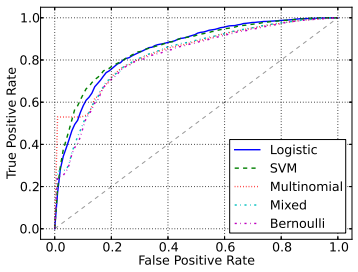
4.3 Analysis of Logistic Regression

We focus on logistic regression to further understand the classification results, since it provides us with coefficients for the movies and confidence in the gender inference. We note that a similar analysis can be done using SVM, which we omit for brevity.

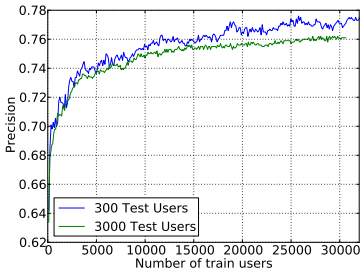
Effect of training set size. Since we use 10-fold cross validation, our training set is large relative to the evaluation set. We use the Flixster data to assess the effect that the number of users in the training set size has on the inference accuracy. In addition to the 10-fold cross validation giving 3000 users in the evaluation set, we performed a 100-fold cross validation using a 300-user evaluation set. Additionally, we



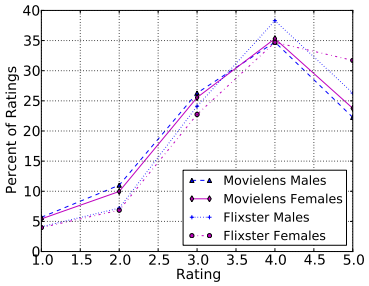
(a) Flixster ROC



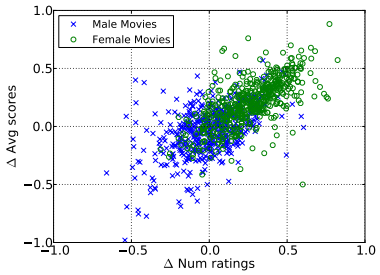
(b) Movielens ROC



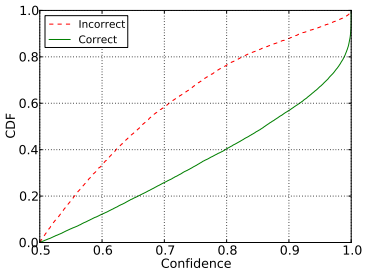
(c) Flixster precision with increasing training set size



(d) Distribution of ratings



(e) Flixster top 500 correlated movies



(f) Flixster CDF of the confidence

Figure 2: Results of gender classification

incrementally increased the training set, starting from 100 users and adding 100 more users on each iteration.

Figure 2c plots the precision of the logistic regression inference on Flixster for the two evaluation set sizes. The figure shows that for both sizes, roughly 300 users in the training set are sufficient for the algorithm to reach above 70% precision, while 5000 users in the training set reaches a precision above 74%. This indicates that a relatively small number of users are sufficient for training.

Movie-Gender Correlation. The coefficients computed by logistic regression expose movies that are most correlated with males and females. Table 3 lists the top 10 movies correlated with each gender for Flixster; similar observations as the ones below hold for Movielens. The movies are ordered based on their average rank across the 10-folds. We use average rank since the coefficients can vary significantly between folds, but the order of movies does not. The top gender correlated movies are quite different depending on whether X or $\tilde{X}$ is used as input. For example, out of the top 100 most female and male correlated movies, only 35 are the same for males across the two inputs, and 27 are the same for females; the comparison yielded a Jaccard distance of 0.19 an 0.16, respectively. We saw that many of the movies in both datasets align with the stereotype that action and horror movies are more correlated with males, while drama and romance are more correlated with females. However, gender inference is not straightforward because the majority of popular movies are well liked by both genders.

Table 3 shows that in both datasets some of the top male correlated movies have plots that involve gay males, (such as Latter Days, Beautiful Thing, and Eating Out); we observed the same results when using $\tilde{X}$. The main reason for this is that all of these movies have a relatively small number of ratings, ranging from a few tens to a few hundreds. In this case it is sufficient for a small variance in the rating distributions between genders with respect to the class priors, to make the movie highly correlated with the class.

Female	Male
Broken Bridges	Latter Days
Something the Lord Made	Beautiful Thing
Drunken Master	Birth
Dracula - Dead and Loving It	Eating Out
Young Indiana Jones	Prince of Darkness
Pootie Tang	Mimic
Anne of Green Gables	Show Girls
Another Cinderella Story	Godzilla: Final Wars
The Fox and the Hound 2	Studio 54
Winnie the Pooh	Desperately Seeking Susan

Table 3: Top male and female correlated movies in Flixster

Further, we evaluate whether the distribution of movie ratings are different for males and females, and whether movies that are correlated with a gender tend to have more ratings from users of that gender. Figure 2d shows that the rating distribution is similar for females and males in the two datasets. In Figure 2e each dot corresponds to a movie; the x-axis plots the normalized difference in the number of ratings between females and males, and the y-axis (scaled to -1 to 1) shows the difference in the value of the ratings between the sexes. Green circles in the upper-right portion (blue crosses in lower left) of the plot indicate that the pair of features (captured on the two axes) can explain the highly female (male) correlated movies, respectively. While this pair of features explains some of the observed gender correlation, they are not nearly sufficient to explain all of it since more than half of the gender correlated movies lie in the middle of this plot.

Confidence in classification. Finally, the confidence value of the classifier is the obstacle that an obfuscation mechanism needs to overcome when trying to hide the gender from the classifier. The higher the confidence of the classifier in

its prediction, the more effort the obfuscation method needs to apply, possibly increasing the impact on the recommendations. Therefore, we evaluate whether the classifier has different confidence values when it outputs a correct or incorrect classification. Figure 2f plots the CDF of the confidence value for correct and incorrect classifications, showing that the confidence is higher when the classification is correct, with a median confidence for incorrect classifications of 0.65, while for correct classification it is 0.85. Moreover, nearly 20% of correct classifications have a confidence of 1.0, which holds for less than 1% of incorrect classifications.

5. GENDER OBFUSCATION

The obfuscation mechanism takes as input a user i's rating profile $\mathcal{H}_i$, a parameter k that represents the number of permitted alterations, and information from the training set to output an altered rating profile $\mathcal{H}'_i$ such that it is hard to infer the gender of the user while minimally impacting the quality of recommendations received. In general, such a mechanism can alter $\mathcal{H}_i$ by adding, deleting or changing movie ratings. We focus on the setting in which the obfuscation mechanism is only allowed to add k movie ratings, since deleting movies is impractical in most services and changing ratings is more suspicious than adding ratings. Because users have different numbers of movies rated in their profiles (and some may have a small number), we do not use a fixed number k but rather we add a number that corresponds to a given percentage of movies in a user's rating profile. In order to add movies into a user's profile, the obfuscation mechanism needs to make two non-trivial decisions:

- Which movies should be added?
- What should be the rating assigned to each movie?

We refer to these added movie ratings as *extra ratings*. We note that the rating values assigned are not "noise" but have some useful value. For example, if this rating corresponds to the average rating over all users, or the predicted rating (using matrix factorization) for a specific user, then the rating value is a reasonable predictor of how the user may have rated had he watched the movie. In this work, we do not aim to provide an exhaustive list of obfuscation mechanisms, instead our goal is to design mechanisms informed by observations from our gender inference study (Section 4).

5.1 Obfuscation mechanisms

To simplify the discussion, we first assume that the obfuscation mechanisms have full access to the training dataset, and can use it to derive information for selecting movies and ratings to add. Later in this section, we amend the assumption of full access to the dataset.

Movie selection. We design three intuitive strategies for selecting movies. Each strategy takes as input a set of movies $\mathcal{S}_i$ rated by the user i, a the number of movies k to be added that corresponds to p% of i's existing profile, and ordered lists L_M and L_F of male and female correlated movies, respectively, and outputs an altered set of movies $\mathcal{S}'_i$, where $\mathcal{S}_i \subseteq \mathcal{S}'_i$. The lists L_M and L_F are stored in decreasing order of the value of a scoring function $w : L_M \cup L_F \rightarrow \mathbb{R}$ where $w(j)$ indicates how strongly correlated a movie $j \in L_M \cup L_F$ is with the associated gender. A concrete example of the scoring function is to set $w(j) = \beta_j$, where β_j is the coefficient of movie j obtained by learning a logistic regression model from the training dataset. We will use this instantiation of the scoring function in our evaluation. Additionally, we assume that $k < \min(|L_M|, |L_F|) - |\mathcal{S}_i|$ and $L_M \cap L_F = \emptyset$.

The movie selection process is as follows. For a given female (or, male) user i, we initialize $\mathcal{S}'_i = \mathcal{S}_i$. Each strategy repeatedly picks a movie j from L_M (or, L_F), and if $j \notin \mathcal{S}'_i$ it adds j to $\mathcal{S}'_i$, until k movies have been added. The set $\mathcal{S}'_i$ is the desired output. The three strategies differ in how a movie is picked from the ordered lists of movies.

1. **Random Strategy.** For a given female (male) user i, pick a movie j uniformly at random from the list corresponding to the opposite gender L_M (L_F), irrespective of the score of the movie.
2. **Sampled Strategy.** Sample a movie based on the distribution of the scores associated with the movies in the list corresponding to the opposite gender. For instance, if there are three movies j_1, j_2, j_3 in L_M with scores $0.5, 0.3, 0.2$, respectively, then j_1 will be picked with probability 0.5 and so on.
3. **Greedy Strategy.** Pick the movie with the highest score in the list corresponding to the opposite gender.

Rating assignment. In Section 4.2, we made a key observation that the binary event of including or excluding a movie in a profile (indicating watched or not) was a signal for gender inference nearly as strong as the ratings. Given that, we aim to assign ratings to the extra movies that have a low impact on the recommendations provided to a user. We propose and evaluate two rating assignments:

1. **Average movie rating.** The obfuscation mechanism uses the available training data to compute the average rating for all movies $j \in \mathcal{S}'_i - \mathcal{S}_i$ and add them to user i's altered rating profile $\mathcal{H}'_i$.
2. **Predicted rating.** The obfuscation mechanism computes the latent factors of movies by performing matrix factorization on the training dataset, and uses those to predict a user's ratings. The predicted ratings for all movies $j \in \mathcal{S}'_i - \mathcal{S}_i$ are added to $\mathcal{H}'_i$.

Access to Dataset. Earlier we assumed the obfuscation mechanism had unrestricted access to the training set. We point out now that our mechanisms described above require access only to the following quantities: (a) for movie selection: ordered lists of male and female correlated movies, and (b) for rating assignment: average movie ratings, and movie latent factors to predict user movie ratings. Note that this information can be found from publicly available datasets, such as the Netflix Prize dataset[2]. Assuming that users in such public datasets are statistically similar overall to those in a particular recommender systems, then we no longer need the assumption of access to the training set.

5.2 Evaluating the obfuscation mechanisms

We evaluate all the permutations of movie selection and rating assignment strategies proposed above. We evaluate values of k corresponding to 1%, 5% and 10% $|\mathcal{S}_i|$ for each user i. The movie scores in lists L_M and L_F are set to the corresponding logistic regression coefficients.

Impact on privacy. We capture the privacy gain that obfuscation brings via the reduced performance in gender inference. Table 4 shows the accuracy of inference for all three movie selection strategies (i.e., random, sampled and

[2]`http://www.netflixprize.com/index`

	Classifier	Strategy	Accuracy with extra ratings			
			0%	1%	5%	10%
Flixster	Logistic Regression	Random	76.5	65.8	46.2	28.5
		Sampled	76.5	60.8	36.6	19.6
		Greedy	76.5	15	1.7	0.1
	Multinomial	Random	71.5	69.3	67	63.5
		Sampled	71.5	68.6	66	61.1
		Greedy	71.5	62	54.3	42.1
Movielens	Logistic Regression	Random	80.2	77.6	71.5	61.1
		Sampled	80.2	75.2	58.6	35.5
		Greedy	80.2	57.7	17.3	2.5
	Multinomial	Random	76.4	75.1	72.9	70.1
		Sampled	76.4	74.9	72.3	68.4
		Greedy	76.4	72.3	66.6	60.4

Table 4: Accuracy of gender inference for different strategies, when rating assignment is average movie rating

	Classifier	Strategy	Accuracy with extra ratings			
			0%	1%	5%	10%
Flixster	Logistic Regression	Random	76.5	65.4	45.5	27.4
		Sampled	76.5	60.5	35.7	18.3
		Greedy	76.5	15.1	1.5	0.1
	Multinomial	Random	71.5	69.5	67.2	63.8
		Sampled	71.5	68.9	66.3	61.5
		Greedy	71.5	63.3	54.9	42.4
Movielens	Logistic Regression	Random	80.2	76.9	68.9	52.7
		Sampled	80.2	73.9	48.9	24.9
		Greedy	80.2	48.4	7.2	0.6
	Multinomial	Random	76.4	74.5	71.8	67.9
		Sampled	76.4	74.3	70.5	65.9
		Greedy	76.4	71.1	64.1	57.3

Table 5: Accuracy of gender inference for different strategies, when rating assignment is users' predicted ratings

greedy) when the rating assigned is the average movie rating. The accuracy is computed using 10 fold cross validation, where the model is trained on unadulterated data, and tested on obfuscated data.

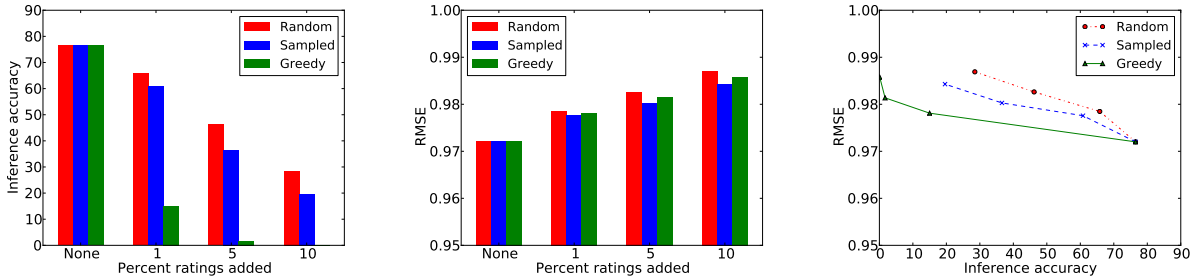
Since the accuracy of inference is the highest for the logistic regression classifier, it would be the natural choice as the inference mechanism for a recommender system. Figures 3a and 3d show the drop in inference accuracy for adding noisy ratings for the two datasets. On adding just 1% extra ratings using the greedy strategy, the accuracy drops to 15% (that is an 80% decrease) and with 10% extra ratings the accuracy is close to zero for the Flixster dataset, as compared with the accuracy of 76.5% on the unadulterated data. Therefore, if the obfuscation mechanism selects movies according to the greedy strategy, adding a small number of movies is sufficient to obfuscate gender. Even when the movies are chosen using the random strategy (which ignores movie scores and hence, the logistic regression coefficients), just 10% additional movies correlated with the opposite gender are sufficient to decrease the accuracy of gender inference by 63% (from 76.5% to 28.5% accuracy). Similar trends are observed for the Movielens dataset.

Our obfuscation mechanism above is using ordered lists that correspond well to the inference mechanism's notion of male or female correlated movies. However, in general, the obfuscation mechanism does not know which inference algorithm is used and thus lists such as L_M and L_F may have a weaker match to such a notion interior to the inference algorithm. We evaluate our obfuscation under such a scenario, with Multinomial Naïve Bayes and SVM classifiers. Our obfuscation still performs well as we see in Table 4, the inference accuracy of the Multinomial classifier drops from 71% to 42.1% for Flixster, and from 76% to 60% for the Movielens dataset (with 10% extra ratings and the greedy strategy). Obfuscation results in a similar decrease in gender inference accuracy of SVMs, results omitted for brevity.

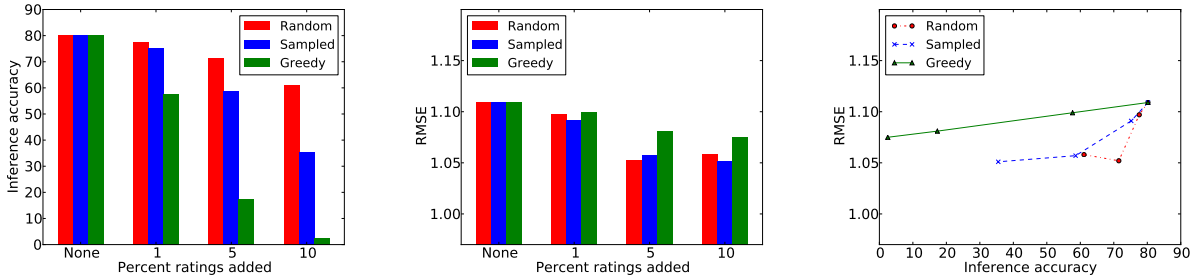
Impact on recommendations. Next, we evaluate the impact on the recommendation quality that the user will observe if she obfuscates her gender. We measure this impact by computing the RMSE of matrix factorization on a held-out test set of 10 ratings for each user. Again, we perform 10 fold cross validation, where the data for users in 9 folds is unadulterated, and one of the folds has users with additional noisy ratings. That is, we use $\mathcal{H}'$ for a tenth of the users, and $\mathcal{H}$ for the rest. This is equivalent to evaluating the change in RMSE for 10% of the users in the system who obfuscate their gender. Figures 3b and 3e show the change in RMSE due to obfuscation for Flixster and Movielens, respectively, when the ratings added were the same as in Table 4. Overall, we see that obfuscation has negligible impact on RMSE. For Flixster, we see that compared to the case of no extra ratings ("none") the RMSE increases with additional ratings, although negligibly. For Movielens, we observe a slight decrease in RMSE with extra ratings. We conjecture that this may occur because by adding extra ratings we increase the density of the original rating matrix which may improve the performance of matrix factorization solutions. Another explanation could be that the extra ratings are not arbitrary, but somewhat meaningful (i.e., the average across all users). The key observation is that for both datasets, the change in RMSE is not significant, a maximum of 0.015 for Flixster (with random strategy and 10% extra ratings), and 0.058 for Movielens (with sampled strategy and 10% extra ratings).

Analyzing privacy-utility tradeoff. We now take a comprehensive look at the privacy-utility tradeoff of the proposed obfuscation, where the desired high privacy corresponds to a low accuracy of gender inference, and a high utility corresponds to a low RMSE which is often used as a proxy for high quality recommendations. Figures 3c and 3f show the privacy (inference accuracy) on the x-axis and utility (RMSE) on the y-axis. Each point on the curve corresponds to the amount of extra ratings, where the rightmost point corresponds to no additional ratings, and the following points moving left are 1%, 5% and 10% extra ratings. For the Flixster dataset, as we move towards higher privacy the utility decreases. As described above, for Movielens as we move towards higher privacy, the utility increases however only slightly. These plots illustrate the clear trend that our obfuscation mechanism can lead to a substantial reduction in gender inference accuracy yet only incurs very small changes to the quality of the recommendations.

Preserving recommendation quality. We now evaluate the tradeoff when the rating assignment corresponds to the "predicted ratings" approach (Section 5.1). The motivation behind this rating assignment is that, in principle, this obfuscation results in no change in RMSE as compared with the RMSE on unaltered data. In other words, there is no tradeoff to be made on the utility front with this choice of rating assignment. Table 5 shows the accuracy of gender inference when this rating assignment is used. The results are similar those in Table 4 where the rating assignment is the average movie rating. For the Movielens data, the accuracy of gender inference is slightly lower with predicted ratings;



(a) Inference on obfuscation - Flixster (b) RMSE on obfuscation - Flixster (c) RMSE Inference tradeoff - Flixster



(d) Inference on obfuscation - Movielens (e) RMSE on obfuscation - Movielens (f) RMSE Inference tradeoff - Movielens

Figure 3: Effect of obfuscation on inference and recommendations

for example, for the greedy strategy with 1% extra ratings, the accuracy of the logistic regression classifier reduces from 57.7% to 48.4% - and this benefit comes without sacrificing the quality of recommendations.

In conclusion, our experimental evaluation shows that with small amount of additional ratings, it is possible to protect a user's gender by obfuscation, with an insignificant change to the quality of recommendations received by the user.

6. CONCLUSION

In this work we show that a user's rating profile alone can be used to infer her gender with high accuracy. Given a relatively small training set, our inference algorithms correctly predict the gender of users with a precision of 70%-80%. We use the insights from inferring gender to design obfuscation mechanisms that add ratings to a user's profile with the goal of making it hard to infer the user's gender, while posing a minimal impact on recommendation quality. We evaluate the tradeoff in the privacy and utility for different obfuscation mechanisms and show that just 1% additional ratings to a user's profile decreases the inference accuracy by 80%.

Although the focus of this paper is on a relatively simple binary inference of the gender, it raises a red flag regarding the possibility to infer private information about users based on the apparently non-revealing act of rating items for purpose of recommendations, unlike the more explicit actions performed in social networks. We plan to further study the accuracy of more sensitive private information in future work.

Acknowledgments. The authors would like to thank Mohsen Jamali for sharing the Flixster dataset.

7. REFERENCES

[1] G. Adomavicius and J. Zhang. On the stability of recommendation algorithms. In *RecSys*, 2010.

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

[3] Z. Cheng and N. Hurley. Effective diverse and obfuscated attacks on model-based recommender systems. In *RecSys*, 2009.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer-Verlag, 2001.

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

[6] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[7] A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In *AAAI*, 1998.

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

[9] 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.

[10] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre. Recommender systems: Attack types and strategies. In *AAAI*, 2005.

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

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

[13] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *KDD*, 2008.

[14] V. Toubiana, L. Subramanian, and H. Nissenbaum. Trackmenot: Enhancing the privacy of web search. *CoRR*, abs/1109.4677, 2011.

[15] S. Ye, F. Wu, R. Pandey, and H. Chen. Noise injection for search privacy protection. In *CSE*, 2009.