Poster: Privacy-Aware Publishing of Netflix Data

Brian Thompson
Dept. of Computer Science
Rutgers University
Piscataway, NJ
bthom@cs.rutgers.edu

Chih-Cheng Chang
Dept. of Computer Science
Rutgers University
Piscataway, NJ
genius@cs.rutgers.edu

Hui (Wendy) Wang
Dept. of Computer Science
Stevens Institute of Technology
Hoboken, NJ
hwang@cs.stevens.edu

Danfeng Yao
Dept. of Computer Science
Rutgers University
Piscataway, NJ
danfeng@cs.rutgers.edu

***Abstract*—To seek better prediction techniques, data owners of recommender systems such as Netflix sometimes make their customers' reviews available to the public, which raises serious privacy concerns. With only a small amount of knowledge about individuals in a recommender system, an adversary may be able to re-identify users and consequently determine their item ratings. In this work, we present a robust and efficient anonymization algorithm for publishing recommendation datasets, *Predictive Anonymization*, that gives desired privacy guarantees without significantly affecting prediction accuracy.**

I. Introduction

Netflix, the world's largest online DVD rental service, recently announced a million-dollar *Netflix Prize* for improving their movie recommendation algorithm. To aid contestants, Netflix released a dataset containing around 100 million movie ratings for 500,000 Netflix subscribers. As the dataset contains users' private preferences to the movies, Netflix replaces names with arbitrary ID numbers to protect their privacy. However, this naively anonymized data suffers from re-identification attacks as recently demonstrated [1].

Yet, growing trends towards openness in data sharing are not only inevitable, but essential to the technological growth of our society. They open doors to scientific research in fields ranging from social psychology to biomedicine, enable the development of products that contribute to the convenience of modern living, and pave the way for a new generation of goods and services that provide valuable societal benefits. In this work, we investigate the feasibility of preserving the privacy of individuals while maximizing the utility that can be gained from releasing large recommender databases to the public.

Although privacy preservation in data publishing has been studied extensively over the last decade, most of these techniques are designed for relational databases or general unlabeled graphs, and are not directly applicable to recommender systems, which we represent as labeled bipartite graphs.

Most importantly, none of the existing work has effectively addressed the impact that sparsity has on anonymization: it increases the probability that de-anonymization succeeds, and increases the difficulty of designing anonymization schemes that provide acceptable predication accuracy. Unfortunately, existing anonymization algorithms are not effective when applied to sparse datasets, which includes most real-world recommender systems. In comparison, we develop a general and efficient approach, *Predictive Anonymization*, that preserves both user privacy and data utility.

Our main idea is that before anonymization, we pad the null entries to reduce data sparsity by performing a round of prediction. This predict-then-anonymize sequence is able to uncover and leverage the latent interests of users that would otherwise be lost without the pre-processing.

Contributions of our work can be summarized as follows:

- We give privacy and attack models for recommendation databases, including an adaptation of the k-anonymity model for relational databases.
- To combat sparsity and preserve data utility, we develop a novel *predictive anonymization* technique to pad, cluster, and anonymize the recommendation data.
- We perform experiments on the Netflix dataset. Our results show that (1) naive anonymization methods incur high information loss, and (2) our predictive anonymization approach is effective in reducing data sparsity while preserving data utility during anonymization.

II. Model

We model a recommendation database as a *labeled bipartite review graph*, where users and items are represented by nodes, and labeled edges correspond to ratings given to items by users. In this work, we aim to achieve two important privacy goals: *node identification privacy*, that the identities of individuals in the released data are considered sensitive, and *link existence privacy*, that it should not be possible to infer whether a particular user had rated a particular item in the recommender system. We consider two adversary models: a *structure-based attack*, where the adversary has background knowledge of which items a user has rated, and the stronger *label-based attack*, in which additionally the adversary knows the corresponding ratings assigned to those items by the user. Our anonymization algorithms are designed to protect users' privacy against the stronger label-based attack.

To give formal privacy guarantees, we adapt the k-anonymity model for relational databases [2] to the context of recommender systems. We say a user u is k-anonymous if in the released database there are at least $k-1$ other users whose lists of item ratings are identical to those of u.

III. Algorithm

To achieve k-anonymity, we develop an efficient approach, Predictive Anonymization, which consists of three major steps: *padding*, *clustering*, and *homogenization*.

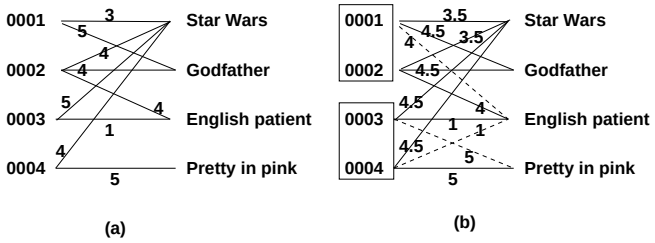

Fig. 1. Review graph (a) before and (b) after Simple Anonymization.

Padding Recommendation data is typically very sparse. Since overlap between any two users is small, cluster-based anonymization techniques may not effectively group similar users, having a devastating effect on prediction accuracy.

We take a novel approach to anonymization by utilizing *singular value decomposition* (SVD) as a pre-processing method before anonymization. All zero-ratings are replaced with predicted values, eliminating the sparsity problem (see Table 1) and significantly improving clustering accuracy.

Dataset	[0]	[1]	[2]	[3]	[4]	[5]
Original	98.8%	0.05%	0.12%	0.33%	0.39%	0.27%
Padded	0%	0.79%	14.1%	46.7%	33.5%	4.89%

Table 1. Distribution of ratings in original and padded data.

Clustering The next goal is to cluster users into anonymization groups, each of size $\geq k$. To achieve this minimum-size requirement, we use the *bounded t-means algorithm* from [3]. When measuring the similarity of two users, we use a *weighted-squared similarity metric*, which gives a squared penalty for large differences in item preferences. In order to efficiently accommodate very large datasets, we employ a sampling technique, the details of which can be found in our full version [5]. Note that throughout the clustering step we use the *padded dataset* to ensure the quality of the clusters.

Homogenization To defend against both the structure-based and label-based attacks, our final step is to *homogenize* the k users in each cluster so that they have identical sets of rated items and corresponding ratings in the anonymized graph. We describe two approaches, each providing different benefits.

In *Simple Anonymization*, we first consider the union of all items rated by users in the cluster. For each item, we take the average rating over all users in the cluster who have rated that item. We then re-assign the edge label from each user to that item to be the average value, adding fake edges as necessary. When homogenization is complete, all users in the cluster have been assigned the same rating for each item (see Fig. 1).

Instead of reverting back to the original data, the *Padded Anonymization* approach homogenizes over the padded data. This results in a complete labeled bipartite graph for the released data, where all users in an anonymization group have the same rating for every item in the database.

Both the Simple and the Padded Anonymization algorithms preserve *node identification privacy* and *link existence privacy* against the stronger *label-based attack* (see Section II). For formal privacy analysis and discussion of further privacy issues such as l-diversity, please refer to our full version [5].

IV. Experiments

We use the entire Netflix dataset for our experiment. The original data contains a total of 480,189 users' ratings on 17,770 movies. The ratings range from 1 to 5, with 0 meaning a rating does not exist. We use the open-source SVD implementation in the Netflix Recommender Framework [4] for padding, and also for prediction when necessary.

We measure prediction accuracy by removing over a million ratings from the dataset, anonymizing the remaining data, and predicting the missing values. We then calculate the *root mean squared error* (RMSE) of the prediction results. In order to clearly quantify the information loss incurred by anonymization, we compare the RMSE values when prediction is performed *before* and *after* anonymization (see Table 2).

Experiment Series	RMSE
Original Data	0.951849
Padded Anonymization (k = 5)	0.95970
Padded Anonymization (k = 50)	0.95871
Simple Anonymization (k = 5)	2.36947
Simple Anonymization (k = 50)	2.3771

Table 2. Accuracy of Predictive Anonymization.

Our results show that *Padded Anonymization* is extremely effective in preserving data quality, with low prediction error (RMSE = 0.959) comparable to that of the non-anonymized data (RMSE = 0.952), even with large values of k. Homogenization on the original data as in the *Simple Anonymization* method gives much higher RMSE, indicating that naive anonymization incurs high information loss even with a small k value. This validates earlier predictions by others [1].

Although preserving prediction accuracy, the padded anonymization method loses some properties of the original data. For example, the released data cannot support statistical queries such as percentage of users who have rated a particular movie. This underlines an intrinsic tradeoff between user privacy and data utility, which is a subject of our future work.

V. Conclusion

In this work, we showed that utility-preserving anonymization for recommendation data is feasible, if careful padding is performed to reduce data sparsity. We defined privacy and attack models, and developed a practical and efficient *Predictive Anonymization* algorithm that preserves both privacy and utility in the anonymized data.

References

[1] A. Narayanan, V. Shmatikov, "How To Break Anonymity of the Netflix Prize Dataset," S&P, 2008.
[2] P. Samarati, Latanya Sweeney, "Generalizing data to provide anonymity when disclosing information," PODS, 1998.
[3] B. Thompson, D. Yao, "Union-Split Clustering Algorithm and Social Network Anonymization," ASIACCS, 2009.
[4] Netflix Recommender Framework. http://benjamin-meyer.blogspot.com/2006/10/netflix-prize-contest.html
[5] C. Chang, B. Thompson, H. Wang, D. Yao, "Predictive Anonymization: Utility-Preserving Publishing of Sparse Recommendation Data," Rutgers University Technical Report, DCS-TR-647, 2009. http://www.cs.rutgers.edu/research/technical_reports/