Leveraging Uncertainty Quantification for Reducing Data for Recommender Systems

1st Xi Niu, 1st Ruhani Rahman, 3rd Xiangcheng Wu, 5th Zhe Fu, 6th Depeng Xu
University of North Carolina at Charlotte
{xniu2, rrahman3, xwu20, zfu2, dxu7}@charlotte.edu

4th Riyi Qiu
Freddie Mac
qiuriyi@gmail.com

2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386790

***Abstract*—The recent California Consumer Privacy Act (CCPA) requires that personal data shall be limited to what is necessary for business purposes. Business services shall "implement technical safeguards that prohibit re-identification of the consumer to whom the information may pertain". For recommender systems, we believe the legal concepts of limitation and technical safeguard are not specific enough to operationalize in practice. This study makes efforts to map the legislative challenges to practice of reducing personal data. More importantly, we borrowed the notion of uncertainty from the machine learning community, and added it as another aspect of recommendation utility, in addition to recommendation accuracy, to guide the data reduction process. The benefit of using uncertainty is that we have more comprehensive consideration while reducing the personal data. In addition, two major types of uncertainty in machine learning models: aleatoric uncertainty and epistemic uncertainty, helped us formulate two groups of data reduction strategies: within-user and between-user. We conducted a series of analyses regarding uncertainty change and accuracy loss caused by different data reduction strategies. We found that at the aggregate level, data reduction is feasible with certain data reduction strategies. At the individual level, the recommendation utility (both uncertainty and accuracy) loss incurred by data reduction disparately impacts different users — a finding which has implications for fairness and transparency of AI models. Our results reveal the difficulty and intricacy of the data reduction problem in the context of recommender systems.**

***Index Terms*—Personal Data Reduction, Uncertainty, Aleatoric Uncertainty, Epistemic Uncertainty, Recommendation Utility, Recommender Systems, Deep Learning**

I. Introduction

In recent years, there has been tension between extensive data collection to achieve online service quality and personal data protection. In 2018, California lawmakers passed *California Consumer Privacy Act (CCPA)* [1], which states that personalized services only retain customers' data necessary to do business. When it comes to recommender systems, it is more challenging to operationalize the key concept of "limited to what is necessary" from the *Act*.

It is common practice today that recommender systems and search engines collect large amounts of user profiles and interaction activities. It is widely believed that such data is generally necessary for the systems to deliver personalized services. However, what data exactly and how much of them are minimally needed to deliver quality personalized results remain unknown. Some recommender systems researchers believe it is possible that reasonable recommendation utility could be achieved with reduced data of the currently stored user interactions. For example, Biega et al. [2] show that the recommendation accuracy decrease in a recommender system incurred by personal data minimization is not substantial. **All of these existing studies addressed the recommendation utility from the perspective of recommendation accuracy: what a model knows with the reduced data, but not from an angle of what a model does not know: the model uncertainty with the reduced data. We believe understanding what a model does not know after data reduction provides us additional insights about the model's incompetence, which is equally essential for decision makers.**

In recent years, deep learning models have achieved tremendous success in recommender systems. Deep learning models have a natural connection with uncertainty due to their neuron placement and stochastic mechanism. Traditionally, uncertainty is modeled in a probabilistic way. However, we believe the probabilistic modeling, capturing uncertainty in terms of a probability distribution, fails to distinguish two inherently different sources of uncertainty, which are often referred to as aleatoric and epistemic uncertainty [3], [4]. Aleatoric uncertainty refers to the notion of inherent noises and randomness, which may cause variability in the outcome of a deep learning model. As opposed to aleatoric uncertainty, epistemic uncertainty refers to uncertainty caused by lack of knowledge due to insufficient model development process. In other words, it refers to the ignorance of the model, and hence to the epistemic state instead of any underlying random phenomenon. An example of an image classification model can well explain the distinction between aleatoric and epistemic uncertainty. When trained on images of dogs, a model cannot make predictions well on a new dog image with poor image quality. The image quality reflects aleatoric uncertainty, which is case-specific. Meanwhile, the model cannot make a good prediction on a picture of a cat either, even with high image quality, due to lack of knowledge.

We believe such a notion of uncertainty and the distinction between the two sources of uncertainty offer insights on when and where a trained model fails. These insights could be leveraged in two important ways in personal data reduction: 1) guiding us to find possible sources of data to reduce, and 2)

This research is supported by National Science Foundation (NSF) (Award #1910696). We are grateful to NSF to make this research possible.

evaluating the corresponding uncertainty trade-offs after the data reduction, as one aspect of measuring recommendation utility loss. The goal of this study is to find data reduction approaches that maintain a balance between user protection and recommendation utility in terms of both accuracy and uncertainty.

Specifically in this paper, we propose two groups of data reduction strategies, based on the distinction between aleatoric and epistemic uncertainty. The first is within-user data reduction, which is to minimize per-user data. Since aleatoric uncertainty is caused by inherent noises and randomness, reducing each user's profile size may change the level of noises and randomness. We will investigate a point or an extent to which such a change is acceptable. The second is between-user data reduction, which is to reduce a type of user samples during the model training. Such reduction is subject to an acceptable level of epistemic uncertainty increase. Epistemic uncertainty is due to a lack of knowledge, which is believed to increase by removing some user types in the training dataset. We characterized users into several types, such as active or inactive users, nice or harsh users, etc. We will investigate for which reduced user type(s) the model can still maintain a reasonable level of epistemic uncertainty. In addition, we believe preserving the average utility conceals substantial details for individual users. Therefore using either data reduction approach, the change of recommendation utility will be investigated at both the aggregate level and the individual user level. We will also compare the impacts of different data reduction strategies on different user characteristics.

Our three major contributions in this paper are:

- Leveraged the notion of uncertainty as one aspect of recommendation utility to guide personal data reduction process.
- Proposed two data reduction strategies: within-user and between-user reductions, based on the distinction between aleatoric and epistemic uncertainty.
- Investigated the data reduction problem from both the aggregate level and the individual user level.

II. Defining the Tasks

Generally speaking, the utility of recommendations is correlated with the amount of the underlying user data. Reduced data would inevitably hurt the recommendation utility. Therefore, trade-offs need to be balanced between utility and such data reduction. First, data must be sufficient to serve the business purpose. Second, data must be restricted to the minimally needed, meaning only necessary data are stored and processed in order to achieve accurate and robust results. Due to the complex nature of users' data collected by recommender systems, both aspects (being sufficient and being necessary) are challenging to operationalize.

The two proposed data reduction approaches are visualized in Figure 1. The left panel (Figure 1(a)) demonstrates the within-user data reduction, where the matrix entries with a tick denote the retained data points and the entries without a tick denote the removed data points. As aleatoric uncertainty is case-specific and has nothing to do with the model training, we will apply this within-user data reduction only on the testing cases. We keep the training cases untouched, so all the reduced data cases will be presented to the same trained model in the testing stage. In contrast, the right panel (Figure 1(b)) shows the between-user data reduction approach. Similarly, the matrix entries with a tick denote the retained data points and the entries without a tick denote the removed data points. Since epistemic uncertainty is caused by a lack of knowledge on some type(s) of cases, we characterize the users into several types, remove a certain portion of one type in the training cases, and then test the epistemic uncertainty change on the testing data with only that reduced type of testing cases. In Figure 1(b), we characterized the users into two types: red and blue. We removed 50% of the training samples for the red type and tested the corresponding change of the epistemic uncertainty only on the red testing cases in the testing data. It is noteworthy that all the blue cases will participate in the training stage to maintain the model's robustness.

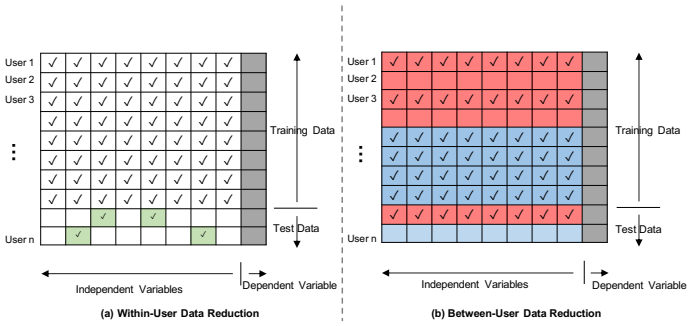

Fig. 1. Within-user and between-user data reduction approaches

Next, we define the goal of each data reduction strategy:

Definition 1: The Goal of Within-User Data Reduction. Let $\hat{\mathbf{r}}_u$ be the predicted ratings using the full data for the user u and $\hat{\mathbf{r}'}_\mathbf{u}$ be the predicted ratings using the reduced data for the user u.

A system satisfies the within-user data reduction goal if it minimizes the amount of individual user data to k data points while achieving recommendation utility compared to using the full data of the user. Recommendation utility comprises two components: accuracy $accu(\hat{\mathbf{r}'}_u)$ and aleatoric uncertainty $ale(\hat{\mathbf{r}'}_u)$. At the aggregate level, the goal could be mathematically represented this way:

$$\begin{aligned} &min\ k,\ |\mathbf{u}| = k,\ \forall u \in \mathcal{U},\ s.t. \\ &\mathbb{E}_\mathcal{U}(accu(\hat{\mathbf{r}'}_u)) - \mathbb{E}_\mathcal{U}(accu(\hat{\mathbf{r}}_u)) \leq \lambda_1\ and \\ &\mathbb{E}_\mathcal{U}(ale(\hat{\mathbf{r}'}_u)) - \mathbb{E}_\mathcal{U}(ale(\hat{\mathbf{r}}_u)) \leq \lambda_2, \end{aligned} \tag{1}$$

where $\mathbb{E}_\mathcal{U}$ is the average operation for all the users. $\lambda_1$and λ_2 are the two thresholds of the recommendation utility change, for accuracy and aleatoric uncertainty respectively.

At the individual level, the goal is:

$$\begin{aligned} &min\ k,\ |\mathbf{u}| = k,\ \forall u \in \mathcal{U},\ s.t. \\ &accu(\hat{\mathbf{r}'}_u) - accu(\hat{\mathbf{r}}_u) \leq \lambda_1\ and\ ale(\hat{\mathbf{r}'}_u) - ale(\hat{\mathbf{r}}_u) \leq \lambda_2. \end{aligned} \tag{2}$$

Definition 2: The Goal of Between-User Data Reduction. A system satisfies the between-user data reduction goal if it

reduces the percentage of training user data of a particular type to p while achieving recommendation utility compared to using the full training data of that user type. Recommendation utility comprises two components: accuracy $accu(\hat{\mathbf{r}'}_u)$ and epistemic uncertainty $epi(\hat{\mathbf{r}'}_u)$.

At the aggregate level, the goal could be mathematically represented this way:

$$\begin{aligned} &min\ p,\ |\mathcal{U}_{\mathcal{R}}| = p|\mathcal{U}|,\ \forall u \in \mathcal{U}_{\mathcal{R}},\ s.t. \\ &\mathbb{E}_{\mathcal{U}_{\mathcal{R}}}(accu(\hat{\mathbf{r}'}_u)) - \mathbb{E}_{\mathcal{U}_{\mathcal{R}}}(accu(\hat{\mathbf{r}}_u)) \leq \lambda_3\ and \\ &\mathbb{E}_{\mathcal{U}_{\mathcal{R}}}(epi(\hat{\mathbf{r}'}_u)) - \mathbb{E}_{\mathcal{U}_{\mathcal{R}}}(epi(\hat{\mathbf{r}}_u)) \leq \lambda_4, \end{aligned} \tag{3}$$

where $\mathbb{E}_{\mathcal{U}_{\mathcal{R}}}$ is the average operation for all the users. λ_3 and λ_4 are the thresholds of the recommendation utility change.

At the individual level, the goal is:

$$\begin{aligned} &min\ p,\ |\mathcal{U}_R| = p|\mathcal{U}|,\ \forall u \in \mathcal{U}_R,\ s.t. \\ &accu(\hat{\mathbf{r}'}_u) - accu(\hat{\mathbf{r}}_u) \leq \lambda_3\ and\ epi(\hat{\mathbf{r}'}_u) - epi(\hat{\mathbf{r}}_u) \leq \lambda_4. \end{aligned} \tag{4}$$

III. Experiment Setup

With the defined data reduction goals, we did a series of experiments with two deep learning models and two datasets, to investigate whether the goals could be achieved under varying conditions.

A. Preliminaries: Estimating Aleatoric and Epistemic Uncertainty

Without loss of generalizability, in literature, the recommendation task for a deep learning model could be regarded as a prediction task [5]–[8], predicting whether an item is of interest to a user. In a deep learning prediction model, the network output for the training sample $\mathbf{u_i}$ is $\mathbf{h}_i$, which is then passed into a dense output layer with weight $\mathbf{W_o}$ and then a Softmax layer to predict the probability vector $\mathbf{p}_i$ of user interests in every item:

$$\mathbf{p}_i = Softmax(\mathbf{W_o}\mathbf{h}_i). \tag{5}$$

In the field of computer vision, Bayesian neural network (BNN) methods have been used to capture the uncertainty in deep learning models. Example studies include [9]–[11]. BNN is believed to be robust to overfitting, enable uncertainty estimation, provide more calibrated models, and learn from small datasets [12]. The key idea of BNN is to represent the model weights with some predefined prior distributions and then train the model to learn the probability density of the posteriors.

A series of studies led by Gal [9], [10] developed an approximate inference technique that performed several stochastic forward passes through the model, and estimated the uncertainty by capturing sample mean and variance. The model scaled well to large data and could be adapted to different deep learning models without changing network structures. Gal's research [9], [10] has offered this study computational approaches to estimating aleatoric and epistemic uncertainty in deep learning models.

Estimating Aleatoric Uncertainty. According to the BNN approach proposed by Kendall and Gal in 2017 [9], a noise term $\mathbf{k}$ was added to the output layer weight $\mathbf{W_o}$, and a Gaussian distribution $N(0, \sigma_i^2\mathbf{I})$ was placed for $\mathbf{kh}_i$. The aleatoric uncertainty was represented by the variance σ_i^2 of the Gaussian distribution.

$$\hat{\mathbf{h}}_i = (\mathbf{W_o} + \mathbf{k})\mathbf{h_i} = \mathbf{W_o}\mathbf{h_i} + \mathbf{kh_i}, \tag{6}$$

$$\mathbf{kh}_i \sim N(0, \sigma_i^2\mathbf{I}), \tag{7}$$

where $\mathbf{I}$ stands for an identity matrix. Then, we predict the probability vector of user interests $\mathbf{p}_i$ using the "corrupted" output $\mathbf{h'}_i$, as in the study of Kendall and Gal [9]:

$$\mathbf{p}_i = Softmax(\hat{\mathbf{h}}_i). \tag{8}$$

Since there is no analytical solution to integrate out the Gaussian distribution of the introduced error $\mathbf{kh_i}$ for a normally used cross entropy loss function, Monte Carlo (MC) simulation is used in Kendall and Gal's study in 2017 [9] to approximate the objective. We will briefly introduce it here. Assume that T times Monte Carlo is simulated, the loss function for this added BNN part is:

$$L = \sum_i \log \frac{1}{T} \sum_{t=1}^{T} \exp\left(\hat{h}_{i,t,c} - \log \sum_{c'} \exp \hat{h}_{i,t,c'}\right), \tag{9}$$

where

$$\hat{\mathbf{h}}_{i,t} = \mathbf{W_o}\mathbf{h}_i + \sigma_i^2\epsilon_t, \epsilon_t \sim N(0, \mathbf{I}), \tag{10}$$

where t represents one MC simulation, $\hat{h}_{i,t,c}$ and $\hat{h}_{i,t,c'}$ are the c and c' elements in the logit vector $\hat{\mathbf{h}}_{i,t}$, and L stands for the Bayesian binary cross entropy loss.

This method only performs Bayesian learning at the output layer. The simulation is performed after the calculation of the network output $\mathbf{h}_i$. Therefore it only increases a fraction of the model computing time. It can be readily applied to any deep learning model with corresponding modifications on the output layer and the loss function.

Estimating Epistemic Uncertainty. In order to estimate epistemic uncertainty, dropout is also performed during the testing stage to sample from the approximate posterior distribution. This dropout contains several stochastic forward passes, referred to as Monte Carlo dropout [9]. For a classification task, epistemic uncertainty is captured by the entropy of the predictive probability vector:

$$H(\mathbf{p}_i) = \sum_{c=1}^{C} p_c \log p_c, \tag{11}$$

where p_c is the average prediction of the T times for a class c, calculated as:

$$p_c = \frac{1}{T} \sum_{t=1}^{T} p_{t,c}. \tag{12}$$

B. Datasets, Prediction Tasks, and Base Models

To demonstrate our ideas, we have experimented with two publicly available datasets and two "vanilla" deep learning models to place the Bayesian layer. Due to the large amount of computation related to the Monte Carlo simulation and dropout for uncertainty estimation, we chose the *MovieLens*-1M [13] dataset to balance the computational cost and the

scalability. The preprocessing procedures of the movie ratings followed the procedures in [14]. In addition, we selected the *Amazon Reviews Dataset* [15] as a base dataset, and made a dataset called *AmazonBooks* with similar number of ratings as *MovieLens*-1M in order to compare our experiment results. Some basic statistics of the two datasets are given in Table I. Please note that *AmazonBooks* is sparser than *MovieLens*-1M, providing the possibility to test the uncertainty robustness for datasets with different sparsity levels.

TABLE I
STATISTICS OF THE DATASETS

Datasets	#Users	#Items	#Ratings	Density
MovieLens	6,040	3,416	1,000,206	0.0426
AmazonBooks	8,1933	7,500	948,750	0.0015

The prediction task was set to be the next item (movie/book) prediction, where the first $n-1$ items in chronological order were used as input features (independent variable) and the last item was used as the target variable. BERT4Rec [16] and NextItNet [17] were adopted to be the base models. The former is Transformer-based and the latter is CNN-based, representing two major types of state-of-the-art sequence-based deep learning models. The difference in recommendation accuracy and uncertainty between the two models with contrasting structures will be compared.

C. Data Reduction Strategies

For the within-user data reduction strategy, for each user, we selected a subset of ratings to reduce the size of the user's profile to investigate the change of aleatoric uncertainty and prediction accuracy. The reduction strategies are:

- *Full*. This is based on the full set of ratings for all the users in the testing data. It is the baseline group to compare with other within-user data reduction strategy groups.
- *Random*. This strategy randomly selects k ratings from each user in the testing dataset to reduce the profile size. If ratings of a certain type (e.g., with a high value) are common in the full profile, they are likely to be preserved through this random sampling.
- *Most/Least Recent*. These two strategies select k most recent ratings and k least recent ratings from each user in the testing dataset respectively. The purpose is to examine the impact of rating recency on the aleatoric uncertainty. It is noteworthy that if a user has a long time span of ratings, we would expect the uncertainty to be largely different from the full group.
- *Most/Least Favorite*. These two strategies select the k ratings that have the highest or the lowest values for each user in the test dataset, respectively. This is to investigate the impact of removing the strong likes or strong dislikes in the aleatoric uncertainty. If a user's rating variance is large, the aleatoric uncertainty of these two strategies would be largely different from the full group.
- *Most/Least Rated*. This method uses the entire dataset to calculate the item popularity measured by the number of ratings. For a given user in the testing dataset, we select the k ratings for the k items that have been rated the most or the least often. The purpose is to check the impacts of the item popularity on this user's profile in terms of aleatoric uncertainty.

For the between-user data reduction strategy, we selected a subset of users in the training dataset to provide to a recommender algorithm. The purpose is to reduce the number of users in our training set without sacrificing too much of epistemic uncertainty and prediction accuracy. We used the following data reduction strategies:

- *Full*. This is based on the full set of users. We compare other between-user data reduction strategies against this baseline group.
- *Most/Least Active*. This strategy selects a certain portion p of the most active users or a certain portion p of the least active users into the training data. Active users are defined by their number of ratings. It is to examine whether the reduced active (or inactive) users in the training dataset could be covered well by the rest with an acceptable sacrifice of epistemic uncertainty and prediction accuracy.
- *Nicest/Harshest*. This strategy respectively selects a certain portion p of the nicest users who have the highest average ratings or the harshest users with the lowest average ratings into the training dataset. The purpose is to investigate the impact on the epistemic uncertainty and prediction on the nicest users (or the harshest users).
- *Most/Least Consistent*. This strategy selects a certain portion p of the most (least) consistent users with their ratings having the least (most) variance. If most users' variances are similar, these two reduction strategies would get similar recommendation quality with the *Full* strategy.
- *Most/Least Typical*. This strategy selects a certain portion p of the most typical users or the most atypical users. Being typical and being atypical are defined by a user's average similarity to other users in the whole dataset. The goal is to check whether an atypical user could be represented well by those typical users.

IV. WITHIN-USER DATA REDUCTION RESULTS

This set of experiments is to compare the impacts of different within-user data reduction strategies on aleatoric uncertainty and prediction accuracy, measured as *hit rate at top k*. We compared the recommendation utility of a particular data reduction strategy against if the recommender algorithm saw the full set of user ratings (the *Full* strategy).

A. Aggregate Results

Figure 2 presents the aleatoric uncertainty and hit rate comparisons of various within-user data reduction strategies. For the limited space, we only present the results of the BERT4Rec-based BNN model on the *MovieLens* data. Later, we will briefly describe the result difference for different models and different datasets. The parameter k denotes the number of ratings sampled from each user's profile data and presented into the trained recommender model. For each

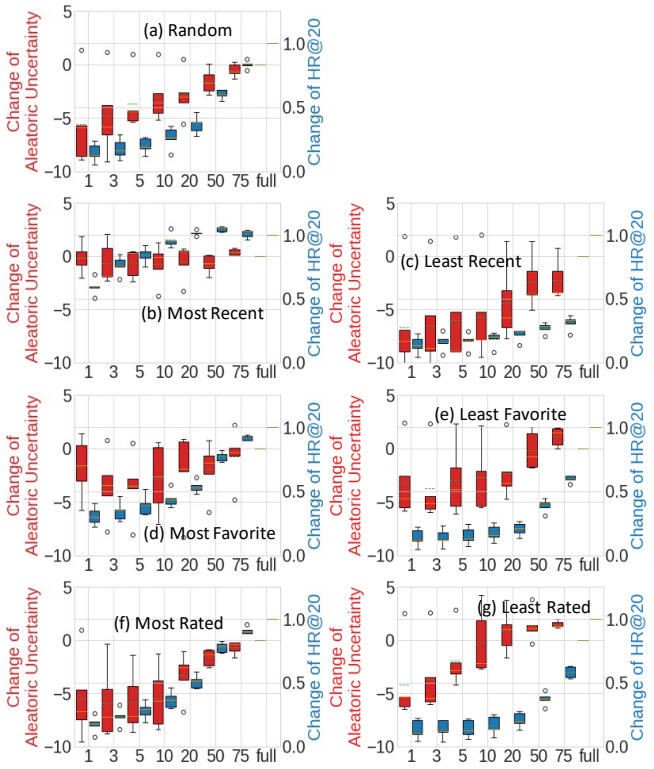

Fig. 2. Within-user data reduction strategies and recommendation utility change

data reduction strategy, we adopted a 5-fold cross validation approach, meaning a BNN model was tested five times. Each boxplot in Figure 2 presents the distribution of the 5 results for the 5 test folds. Since aleatoric uncertainty values do not carry intrinsic meaning and they only have comparative sense in the results. Therefore we used the ratio of the reduced data's aleatoric uncertainty to the full data's to represent the change. In order to scale down those outliers and control all the results in a presentable range, we took the logarithm of such a ratio as the value of the vertical axis for that reduction strategy. For the *hit rate at top k*, we set 20 as the k value, represented as *HR@20*. We used the ratio of reduced data's *HR@20* compared to the full dataset as the change.

In Figure 2, there are two key observations. First, for most within-user data reduction strategies, the aleatoric uncertainty increases as we retain more data from a user, a surprising result than expected. Since aleatoric uncertainty speaks to the user data's intrinsic noises and randomness, the reason could be that more data in fact brings in more data variability, and therefore a higher level of aleatoric uncertainty. From the aleatoric uncertainty perspective, removing within-user data in fact helps recommendation utility. Second, for the *Most Recent* strategy (Figure 2(b)), the aleatoric uncertainty has maintained a comparable level for all the k values with the *Full* strategy. *HR@20* has stabilized to a comparable value with the *Full* strategy when k is 10 and beyond. By "comparable" we mean the difference between that strategy group and the *Full* strategy is not significant at the .05 level for the t-test. The results suggest the usefulness of rating recency for within-data reduction. Even if we only retained a few most recent ratings to predict users' interests, we could still harvest reasonable recommendation utility in terms of both aleatoric uncertainty and prediction accuracy. Recency is robust to the uncertainty change and prediction accuracy loss introduced by aggressively reducing per-user ratings. The implication for the recommender system industry might be the policies toward keeping more recent data without retaining a long history of user activities.

In contrast, for other data reduction strategies, such as *Least Recent*, *Random*, *Most Favorite* and *Least Rated*, we see the aleatoric uncertainty consistently increases as we keep more data for the users in the test dataset. Meanwhile, *HR@20* is steadily increasing too. For these data reduction strategies, t-tests were conducted and each k group is significantly different compared to its *Full* strategy counterpart. That means if we reduce data from each user, we will end up with two conflicting aspects of recommendation utilities: the smaller (better) aleatoric uncertainty but the lower (worse) prediction accuracy. The results suggest the limited usefulness of these data reduction strategies.

B. Comparison of Datasets and Models

We ran the experiments using two base models (BERT4Rec and NextItNet) and two datasets (*MovieLens* and *AmazonBooks*) respectively. Although each dataset and each model demonstrate some nuance and subtlety, we observed similar trends for aleatoric uncertainty change and *HR@20* change for each data reduction strategy. One major difference is that compared to *MovieLens*, *AmazonBooks* is less sensitive to within-user data reduction, as both uncertainty and accuracy changes are smaller. We attribute it to the different sparsity levels of the two datasets. *AmazonBooks* is sparser, meaning averagely speaking, each user has fewer ratings in the dataset. The higher sparsity level makes the dataset more robust to further within-user data reduction. In contrast, the model difference is not as much as the dataset difference. The only major model difference is that BERT4Rec consistently has higher *HR@20* values than NextItNet, meaning better recommendation accuracy.

C. Individual User Results

Figure 2 in the previous section, Section IV.A, showed that the *Most Recent* strategies are able to satisfy a fixed utility change threshold at the aggregate user level. However, we believe such data reduction strategies may disparately impact different individuals. Not every individual is able to maintain a specified recommendation utility. We investigated whether the aleatoric uncertainty change (compared to the *Full* strategy) incurred by within-user data reduction was correlated with different user characteristics. We considered the following four user characteristics (measured using the user's full profile before any data reduction): (1) number of ratings, (2) average

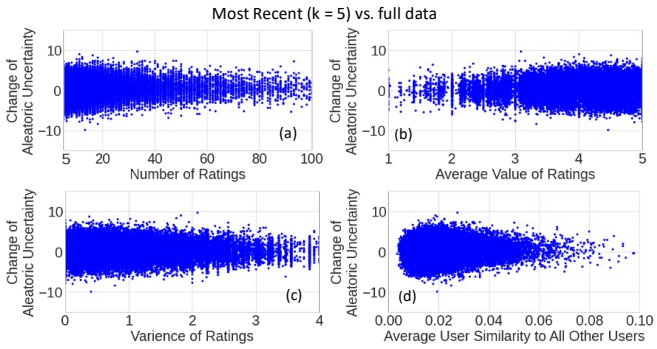

Fig. 3. Aleatoric uncertainty change vs. user characteristics

of the ratings, (3) variance of the ratings, and (4) average similarity to all other users in the whole dataset.

Due to the space limit, we only present the result of the NextItNet-based BNN model on the *AmazonBooks* data. We have obtained similar trends from the BERT4Rec-based BNN model and the *MovieLens* data. Figure 3 shows the aleatoric uncertainty difference (log ratio) between the *Most Recent (k = 5)* strategy and the *Full* strategy. Each user is represented as a data point in each of the four scatterplots.

As in Figure 3(a), as a user's number of ratings increases, the amount of aleatoric uncertainty change keeps reducing. This confirms our expectation that the users with more ratings are more "stable" or robust to within-user data reduction. Active users with more ratings are more likely to be well represented by their most recent 5 ratings. Many users have better (lower) aleatoric uncertainty compared to their full profile data, as we can see more data points showing up in the lower half (below the horizontal axis) of the scatterplot. The improved aleatoric uncertainty means keeping only the most 5 recent ratings may in fact help these users get rid of data noises and randomness.

Figure 3(b) shows the aleatoric uncertainty difference versus the user's average rating value. We observed the users with a lower average rating tend to be more robust to within-user data reduction. On the other hand, users with a higher average rating value are subject to larger aleatoric uncertainty change. The reason is that in this *AmazonBooks* dataset, the average rating is 4.2 out of 5, highly skewed to the higher values. The skewed distribution makes the nicer users less distinguishable than the harsher users. The implication for a recommender system is that for those harsher users, the platform does not need to maintain the same profile size as the nicer users in order to make reliable predictions.

As in Figure 3(c), users with a larger rating variance tend to have smaller aleatoric uncertainty change, suggesting more robustness to within-user data reduction. Variance reflects the degree to which the user tends to distinguish between good and bad items. Those users with a larger rating variance make themselves more robust to within-user data reduction.

Figure 3(d) shows that the users who have high similarity to other users tend to be more robust to within-user data reduction. In contrast, for a user who tends to be different from others, a larger amount of ratings of this user is needed to learn the subtlety of this user's preferences. This reflects the reality that the "non-mainstream" users will be harmed the most when reducing the within-user data.

V. Between-User Data Reduction Results

Between-user data reduction is to reduce the number of users in the training dataset in recommender systems to protect the largest possible number of users. We investigated the extra epistemic uncertainty incurred if we reduced the number of training samples. This set of experiments is to compare the impacts of different between-user data reduction strategies on epistemic uncertainty and prediction accuracy.

We categorized users based on those four user characteristics listed in Section IV.C. They are number of ratings, average of the ratings, variance of the ratings, and average similarity to all other users. For each user characteristic, we used the median value to divide users into two types. Therefore we had eight user types: most/least active users, nicest/harshest users, most/least consistent users, and most/least typical users. These eight user types guided our eight between-user data reduction strategies. For each strategy, we sampled a portion p of that type to retain in the training data, and tested the corresponding epistemic uncertainty on the testing data with only that type of testing cases.

A. Aggregate Results

Figure 4 presents the epistemic uncertainty and hit rate comparisons of various between-user data reduction strategies. Same as the within-user strategy aggregate results, these results are from BERT4Rec-based BNN model on the *MovieLens* data. We will also compare the result differences for different models and different datasets. The sampled percentage p in the X axis of each figure denotes the percentage of retained users of a type in the training data. It is noteworthy that the other type of users was fully retained in the training dataset because the purpose was to check whether this target type, if removed part of its training samples, could be represented well by the rest and the other type. Similar to within-user strategies, for each between-data reduction strategy, we adopted the 5-fold cross validation approach. Each box plot shows the result distribution of epistemic uncertainty change or *HR@20* change for the 5 test folds. Similar to the within-user reduction, we took the log ratio of the reduced data epistemic uncertainty against its full dataset epistemic uncertainty, presented as the value of change of epistemic uncertainty. For the hit rate, we used the ratio of the reduced data's *HR@20* compared to the full dataset as the change.

As in Figure 4(c), for the *Nicest* strategy, we have seen that even the 0% group is able to maintain a comparable epistemic uncertainty with the *Full* strategy (no significant difference at the .05 level for the t-test). HR@20 has no significant difference either. The result indicates that for the nicest users, a recommender system does not have to maintain a full set of users of this type. It is reasonable to claim that as long

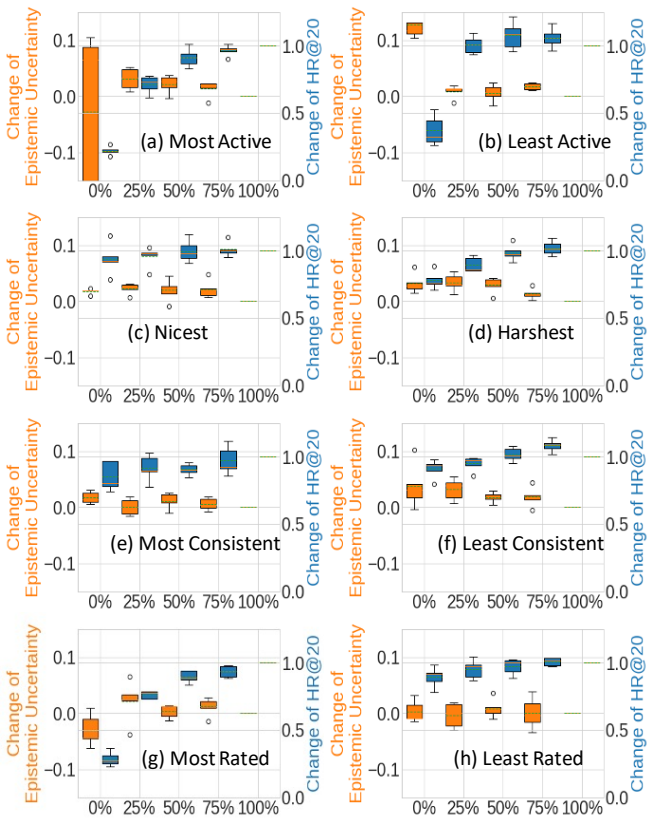

Fig. 4. Between-user data reduction strategies and epistemic uncertainty change

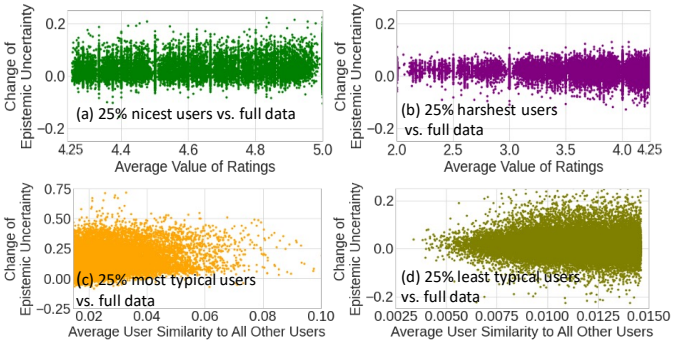

Fig. 5. Epistemic uncertainty change vs. user characteristics

as the model sees a reasonable number of such users, its recommendation utility would not be significantly different from seeing all of such users. In contrast, we cannot make the same claim for the *Harshest* strategy as we can see from Figure 4(d) that as the training samples decrease (from right to left), both the epistemic uncertainty and the HR@20 are getting worse.

Similar to the *Nicest* users, for the *Most Consistent*, we see a relatively comparable epistemic uncertainty and hit rate among different percentage groups (no significant difference in the t-tests). The results mean that these consistent users could be well represented by the inconsistent users if being absent in the training data.

For other strategies, we see the epistemic uncertainty consistently getting worse as the recommender model sees fewer users of that type in the training dataset. All t-tests between a p group and the full data group show significant difference at the .05 level. Meanwhile, *HR@20* is steadily decreasing significantly. These trends suggest for the users of these types, more training samples would help models gain more needed knowledge to make accurate and reliable predictions. Removing training samples is not ideal for these strategies.

B. Comparison of Datasets and Models

Similar to within-user data reduction, for between-user data reduction strategies, we ran the same experiments using two models (BERT4Rec and NextItNet) on two datasets (*MovieLens* and *AmazonBooks*). We find similar trends for epistemic uncertainty change and *HR@20* change for each data reduction strategy. One key observation is that BERT4Rec has seen a relatively "flat" trend of epistemic uncertainty change, suggesting its more robustness to between-user data reduction compared to NextItNet. We attribute it to the Transformers' power, compared to a convolutional structure.

C. Individual User Results

Since the between-user data reduction strategies were guided by the user characteristics, we investigated the impacts of the user characteristics on epistemic uncertainty change. Figure 5 shows two pairs of between-user data reduction strategies versus their corresponding user characteristics: *Nicest (25%)/Harshest (25%)* versus the average of ratings, and *Most Typical (25%)/Least Typical (25%)* versus the average similarity to all other users. We only present the result of the NextItNet-based BNN model on the *AmazonBooks* data. We have similar trends from the BERT4Rec-based BNN model and the *MovieLens* data. From Figure 5, we observe that most data reduction strategies behaved as we expected. The *Harshest (25%)* users represent well the users with a lower average rating as the epistemic uncertainty change is smaller. Similarly, the *Most typical (25%)* users have better robustness for users with a higher similarity to other users, and the *Least typical (25%)* for users with a lower similarity. However, there is one exception. The *Nicest (25%)* users do not necessarily represent better the users with a higher average rating. The possible reason could be the user characteristic variation within the nicest users is larger than that within the harshest users.

VI. Data Reduction and Privacy

In this paper, rather than purely proposing data reduction strategies, we look at the data protection more broadly. We believe the data reduction is related to some computational concepts of the user privacy. We will briefly discuss the relationship. **Identifiability** is a computational concept of privacy. The presence of a unique combination of items in an anonymous user profile poses a de-anonymization risk: if an attacker has the background knowledge that a user has rated these items, they can uniquely identify the user. Inspired

by the work on k-anonymity and related concepts [18], we quantify identifiability through a lower bound on the number of items an attacker would need to know to identify a user. We calculated the identifiable risk of various within-user and between-user data reduction strategies for both the *MovieLens* and *AmazonBooks* datasets. For example, Table II presents the results for between-user data reduction strategies on the *MovieLens* data. The results suggest that all between-user reduction strategies lead to a lower risk of user identifiability compared to that of the full data. The *Most Active* and *Least Consistent* strategies achieved the lowest identifiability risks. For example, an attacker would need to know exactly 4.8 movies that have been rated by a particular user before uniquely identifying this user, under the 25% *Most Active* reduction strategy. This number for the full dataset is only 0.7.

TABLE II
IDENTIFIABILITY FOR USER PROFILES USING DIFFERENT BETWEEN-USER DATA REDUCTION STRATEGIES ON *MovieLens*, AVERAGED ACROSS ALL USERS

Between-User Data Reduction Strategies	Percentage of Users Retained			
	25%	50%	75%	100%
Nicest	4.7	1.9	1.1	0.7
Harshest	4.4	1.8	1.0	0.7
Most Active	4.8	2.0	1.1	0.7
Least Active	4.7	1.8	1.0	0.7
Most Typical	4.6	1.9	1.0	0.7
Least Typical	4.6	1.9	1.0	0.7
Most Consistent	4.4	1.8	1.0	0.7
Least Consistent	4.8	2.0	1.1	0.7

VII. DISCUSSIONS AND CONCLUSION

In this paper, we have mapped the legal concepts of CCPA to practical data reduction operations. We proposed two groups of data reduction strategies, guided by the two sources of uncertainty. The first group focuses on per-user profile size reduction to avoid possible user re-identification while the second focuses on reducing the number of users in the training dataset to protect the largest possible number of users. We evaluated the recommendation utility change in terms of both uncertainty and recommendation accuracy.

We argue that it is possible to achieve reasonable personalization with reduced user data. The recommendation utility incurred by data reduction may not be substantial, but it may disproportionately impact different users — a finding which has implications for fairness and transparency of AI models. We find the usefulness of rating recency for user's profile reduction and argue that recommender systems do not need to maintain a long history of user interactions. We also find that users who are generous in giving high ratings and who are consistent in their ratings could be well represented by other users if their data is absent during the model training. It is observed that data reduction hurts "non-mainstream" users, in particular those who have not used the recommender services often with few ratings and who are atypical compared to the majority users. The "non-mainstream" users will need more data to get recommender services of comparable utility with the "mainstream" users. Some users in fact benefit from some data reduction strategies because those data reductions help them get rid of some data noises or randomness. Our analysis with the user characteristics shows that some variations in individual-level uncertainty change are not well explained by the user characteristics. We will leave the question to future work.

REFERENCES

[1] S. of California. The california consumer privacy act of 2018. [Online]. Available: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

[2] A. J. Biega, P. Potash, H. Daumé, F. Diaz, and M. Finck, "Operationalizing the legal principle of data minimization for personalization," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 399–408.

[3] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural Safety*, vol. 31, no. 2, pp. 105–112, 2009.

[4] S. C. Hora, "Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management," *Reliability Engineering & System Safety*, vol. 54, no. 2-3, pp. 217–223, 1996.

[5] X. Fan and X. Niu, "Implementing and evaluating serendipity in delivering personalized health information," *ACM Transactions on Management Information Systems (TMIS)*, vol. 9, no. 2, pp. 1–19, 2018.

[6] E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "Securebert: A domain-specific language model for cybersecurity," in *International Conference on Security and Privacy in Communication Systems*. Springer, 2022, pp. 39–56.

[7] N. Nur, N. Park, M. Dorodchi, W. Dou, M. J. Mahzoon, X. Niu, and M. L. Maher, "Student network analysis: a novel way to predict delayed graduation in higher education," in *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*. Springer, 2019, pp. 370–382.

[8] X. Niu, C. Lown, and B. M. Hemminger, "Log based analysis of how faceted and text based search interact in a library catalog interface," in *Proceedings of Third Workshop on Human-Computer Interaction and Information Retrieval*. Citeseer, 2009.

[9] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, 2017, pp. 5574–5584.

[10] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, PhD thesis, University of Cambridge, 2016.

[11] A. Siddhant and Z. C. Lipton, "Deep bayesian active learning for natural language processing: Results of a large-scale empirical study," *arXiv preprint arXiv:1808.05697*, 2018.

[12] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," *Advances in Neural Information Processing Systems*, pp. 1019–1027, 2016.

[13] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 1–19, 2015.

[14] R. Devooght and H. Bersini, "Long and short-term recommendations with recurrent neural networks," in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 2017, pp. 13–21.

[15] J. McAuley, C. Targett, J. Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

[16] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019.

[17] F. Yuan, A. Karatzoglou, I. Arapakis, J. M. Jose, and X. He, "A simple convolutional generative network for next item recommendation," *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 582–590, 2019.

[18] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.