Protecting the Anonymity of Online Users through Bayesian Data Synthesis

Matthew J. Schneider*Jingchen Hu†and Shawn Mankad‡

Abstract

User-generated content (UGC) is a modern form of communication that combines multiple types of content such as text, video, audio, images, and other data. However, privacy concerns emerge when online users of popular UGC platforms are identified through a combination of their structured data (e.g., location and name) and textual content (e.g., word choices and writing style). To overcome this privacy problem, we develop a sequential synthesis strategy to generate protected structured data based on a user's adjoining textual content in their UGC. The Bayesian data synthesis model enables platforms to use a single shrinkage parameter to control the privacy level of their released UGC data. Our results show that the sequential synthesis process decreases the probability of identification of a user to an acceptable threshold while maintaining much of the textual content present in the structured covariates. These findings encourage UGC platforms that wish to be known for consumer privacy to protect anonymity of their online users.

*Drexel University, LeBow College of Business, 3220 Market Street, Philadelphia, PA 19104, matt.schneider@drexel.edu

†Vassar College, Box 27, 124 Raymond Ave, Poughkeepsie, NY 12604, United States, jihu@vassar.edu

‡Cornell University, Samuel Curtis Johnson Graduate School of Management, 106 East Ave, Ithaca, NY 14853, spm263@cornell.edu

Keywords: Bayesian data synthesis, disclosure risks, privacy by design, probability of identification, structured data, textual content, user-generated content

1 Introduction

User-generated content (UGC) published on online platforms have an enormous impact on consumer choices and business operations. For example, retailers and service providers analyze the textual content of online reviews to identify operational issues, such as damaged items from shipping, that can be improved in specific states or locations (Laskowski, 2014; Mankad et al., 2016). Moreover, consumers use reviews to guide their purchasing decisions (Jiménez and Mendoza, 2013). However, the abundance of user-level structured data (e.g., location, age, and gender) adjoined with textual content on these platforms also leads to data privacy concerns. While authors of UGC often assume that the posts they contribute are anonymous, others may seek to identify the author of the content (Beigi and Liu, 2020). In one case, an online Yelp user "Birdzeye B" wrote a one-star review about a law firm in the state of California. The law firm identified and sued Ava Bird, resulting in a judgment for over half a million dollars (Porter, 2019). In a second case, an accountant in California confirmed that a Yelp user "Alex M" was a known female customer through an appeals court decision (Jones, 2017).

Thus far, platforms have put the onus on users to protect themselves. For example, Yelp's terms of service states that "You assume all risks associated with Your Content, including anyone's reliance on its quality, accuracy, or reliability, or any disclosure by you of information in Your Content that makes you personally identifiable." Yet, the increasing prevalence of such privacy-related incidents can translate into negative brand effects (Kumar and Reinartz, 2018) or lead to regulation (Campbell et al., 2015; Wieringa et al., 2021). For example, in the United States, the Consumer Reviews Fairness Act was enacted to help preserve the anonymity and integrity of UGC, and the General Data Protection Regulation (GDPR) in Europe has made privacy and data permissions central to social media platform operations.[1] Moreover, platforms that protect anonymity become known for consumer privacy, which leads to a positive brand experience and competitive advantage (Goldfarb and Tucker, 2013; Martin and Murphy, 2017).

Our work attempts to provide a realistic solution to this privacy problem through the generation of partially synthetic data (Little, 1993), where the structured data covariates, such as location, are synthesized from models estimated on the confidential data, while the text remains unchanged. This approach simultaneously protects the data privacy of users while preserving the data utility of the actual post for the general public. A few synthetic data approaches to protection location information have been proposed and evaluated. Burgette and Reiter (2013); Paiva et al. (2014); Quick et al. (2015); Hu and Savitsky (2018) proposed synthesis models based on spatial modeling and directly synthesize variables containing geographical information. Non-spatial synthesis models have also be used and investigated, such as classification and regression trees (CART) (Wang and Reiter, 2012; Hu and Savitsky, 2018; Drechsler and Hu, forthcoming) and Dirichlet Process mixtures of products of multinomials (DPMPM) (Hu and Savitsky, 2018; Drechsler and Hu, forthcoming), which create synthetic geographical variables. Moreover, Quick et al. (2018) used a differential smoothing approach to synthesizing non-geographical information to create synthetic data for public release. To the best of our knowledge, no approach has been proposed

[1] See, for example, https://storytagger.com/resources/gdpr-trust-and-user-generated-content-at-work/

and evaluated for UGC data, where both structured data covariates such as location and textual data are present. Moreover, the aforementioned methods are either highly computationally intensive (e.g. Paiva et al. (2014); Quick et al. (2015, 2018)) or not extendable to textual data in a straightforward manner (e.g. the CART and DPMPM synthesizers in Wang and Reiter (2012); Hu and Savitsky (2018); Drechsler and Hu (forthcoming)).

To tackle the challenge of protecting UGC containing both geographical covariate and textual data in a computationally feasible manner, we propose a data protection model using a Bayesian generalized linear synthesizer for each structured data covariate in a document, following the sequential synthesis strategy for synthesizing more than one variable (Kinney et al., 2011). We define a document as a row of UGC (e.g., one online review or UGC posting) which includes structured data and textual content written by a single user. Specifically, our approach starts with modeling and synthesizing the first covariate conditional on the text. Next, the second covariate is modeled conditional on the text and confidential values of the first covariate, while synthetic values for the second covariate are simulated conditional on the synthetic values of the first covariate from its own synthesis step. The process continues until all sensitive covariates are synthesized. The UGC platform can then release the synthetic structured data in place of the confidential one, together with the un-synthesized text.

When modeling the structured covariates using text, we solve a high dimensionality issue of having more text covariates (e.g., bag of words) than observations by encouraging shrinkage through a Laplace prior distribution on the variance of the text coefficients (Tibshirani, 1996). We also incorporate a privacy-preserving hyperparameter, $\lambda > 0$ as the inverse scale parameter of a Laplace distribution. The Laplace distribution is the prior distribution on the variance of the text coefficients, where the coefficients are pushed towards zero (i.e., words are not informative of a document's structured data) when parameter λ is large. Sparsity in the estimated coefficients is particularly useful when words are overly informative to each structured data covariate (e.g., a specific user in a small area) and compromising user privacy. In other words, our data protection model scales the text information present in the protected structured data from full information (unprotected data) to no information (random sampling) formally through the Laplace prior. This strategy allows the UGC platform to choose the trade-off between data utility and data privacy with a single parameter.

Although our synthetic data approach does not achieve pure differential privacy, a formal privacy guarantee (Dwork et al., 2006), it is related to empirical differential privacy (Schneider and Abowd (2015)) through the privacy parameter λ which scales the textual coefficients in the data synthesis model from fully informative to no information (zero) as λ increases. A high value of λ is likely proportional to lower values of ϵ in empirical differential privacy relating to the bound in the posterior odds ratio of the protected structured data. Achieving pure differential privacy (or empirical differential privacy) for all the UGC is challenging because differential privacy is based on the removal or inclusion or change of a row of data which contains potentially a large number of textual variables in a coherent order and paragraph structure (our application contains over 6,000 textual variables). Krishna et al. (2021) proposed ADePT to release short textual phrases (without adjoining structured data) with differential privacy[2] and Feyisetan et al. (2019) used differential privacy to replace words with semantically similar words, neither of which is suitable for our task of protecting structured data and textual data. Releasing differentially private structured data is

[2] For example, "what are the flights on january first 1992 from boston to san francisco" is altered to "what are the flights on thursday from dallas to san francisco."

also possible (for example, PrivBayes by Zhang et al. (2017)), but the attachment of unaltered text to it can include any information in the free text including the confidential structured data itself. Expansion of the textual data from a large number of textual variables to phrases can result in at least tens of millions of textual variables which far exceeds the number of reviews on a typical UGC platform.

We evaluate our data protection model by analyzing the trade-off between data utility and disclosure risk (Duncan and Lambert, 1986). We measure the relative differences in cross tabulations of categorical covariates between the confidential and the synthetic, as our global utility evaluation. To evaluate analysis-specific data utility which closely mimics how companies analyze their online reviews to direct future resources and improve their operations, we use the Structural Topic Model of Roberts et al. (2014), which models the amount of discussion along different topics as a function of the structured data. Specifically, we quantify how well the relationship between structured covariates and discussion of topics is preserved as the structured covariates are protected. To measure the level of privacy protection of users offered by the synthetic UGC, we focus on evaluating its disclosure risks. Specifically, we calculate the probability of identification (Drechsler and Reiter, 2008; Reiter and Mitra, 2009; Hu, 2019) based on a two-stage identification model. The main idea is that the data intruder (the party interested in identifying the UGC author) filters their own external database by first matching on structured covariates, followed by matching on text data in the second stage.

To demonstrate our methodology, we collect online reviews from Consumer Affairs, a popular online review platform with over one million reviews of over 3,000 brands. Our data focuses on a single financial services vendor that offers insurance products in the United States. We find that that the sequential synthesis process decreases the probability of identification of a user to an acceptable threshold while maintaining much of the textual content present in the structured covariates. Detailed results and discussion are given in Section 4.

The remainder of the paper is organized as follows. Section 2 introduces our Bayesian generalized linear model synthesizer with a shrinkage prior. We then describe our utility measures, global and analysis-specific, in Section 3, as well as our measures of disclosure risks. Section 4 is devoted for our real data application of synthesizing a sample from Consumer Affairs to demonstrate the effectiveness of our proposed synthesis model, which preserves a high level of utility while providing an acceptable level of privacy protection (i.e. sufficiently low disclosure risks). We also show how our privacy parameter λ in our synthesis model is able to tune the trade-off between utility and risks for UGC with textual data. We end with Section 5 with a few concluding remarks.

2 Bayesian GLM synthesizer with shrinkage prior

Our data protection approach is through the generation of partially synthetic data (Little, 1993), where a subset (or the full set) of the structured data covariates, such as state and rating, is synthesized, while textural content remains un-synthesized. Our proposed data synthesizer is a Bayesian generalized linear model for each structured data covariate to be synthesized, conditioning on the textual content and additional structured data covariates, if available.

Assume each document i contains textual data, denoted as $\boldsymbol{x}_i$, and J structured data categorical covariates of each document, denoted as a vector $\boldsymbol{y}_i$. Textual data $\boldsymbol{x}_i$ is un-synthesized and used as covariates in the synthesis models for $\boldsymbol{y}_i$. When $J = 1$, structured data $\boldsymbol{y}_i$ reduces to y_i. To work with the more realistic scenario where multiple structured data categorical covariates are available and a subset of

r of them is to be synthesized for protection, we assume $J > 1$. We use $\boldsymbol{X}$ and $\boldsymbol{Y}$ to denote the textual data and structured data for the entire dataset containing n documents, respectively, and $\boldsymbol{D} = [\boldsymbol{X}, \boldsymbol{Y}]$ to denote the confidential dataset.

We model and synthesize r ($r \leq J$) categorical covariates in a sequential manner (Kinney et al., 2011). The general approach is to start with modeling and synthesizing the first covariate. Next, model the second covariate conditional on the *confidential* values of the first covariate, and synthesize the second covariate conditional on the *synthetic* values of the first covariate from its own synthesis step. The process continues till all r covariates are synthesized.

Without loss of generality, assume we model and synthesize these r categorical covariates in the order from 1 to r. That is, for the first categorical covariate y_{i1}, we use only the textual data $\boldsymbol{x}_i$ as covariates; for the jth covariate y_{ij}, we use categorical covariates $(y_{i1}, y_{i2}, \cdots, y_{i(j-1)})$ in addition to $\boldsymbol{x}_i$ as covariates. For notation simplicity, we use $\tilde{\boldsymbol{x}}_i$ to denote the collection of covariates including $\boldsymbol{x}_i$ and $(y_{i1}, y_{i2}, \cdots, y_{i(j-1)})$, if available.

Assume each y_{ij} takes on values in $\{1, \cdots, K_j\}$, where $K_j \geq 2$ is the number of levels for covariate j. We use a Bayesian generalized linear model (GLM) conditional on the collection of covariates $\tilde{\boldsymbol{x}}_i$ (we note that for first covariate $j = 1$, $\tilde{\boldsymbol{x}}_i$ reduces to x_i). Let π_{ijk_j} denote the probability that document i covariate j belongs to level k_j. Then, the log-odds of each response follows a linear model:

$$\log\left(\frac{\pi_{ijk_j}}{\pi_{ijK_j}}\right) = \tilde{\boldsymbol{x}}_i \boldsymbol{u}_{k_j}, \tag{1}$$

where $\boldsymbol{u}_{k_j}$ is a vector of regression coefficients for covariate j with levels $k_j = 1, \cdots, K_j$ with variance σ^2, and π_{ijk_j} is the probability that document i covariate j belongs to level k_j.

One challenge with textual data is that the number of covariates in the content ("bag of words") can easily become larger than the number of observations (documents). To induce sparsity of textual coefficients, we use independent Laplace priors (with zero mean) for regression coefficients vector $\boldsymbol{u}_{k_j}$. Specifically, we assume coefficients in $\boldsymbol{u}_{k_j}$ are independent with zero mean and constant variance $\sigma^2 = (2/\lambda)^2$ with $\lambda > 0$ according to a Laplace distribution:

$$p(\boldsymbol{u}_{k_j}) = \frac{\lambda}{2} \exp\left(-\lambda |\boldsymbol{u}_{k_j}|\right). \tag{2}$$

We label λ the privacy protection parameter in our synthesizer and note that when λ is large, the variance of the Laplace distribution approaches 0, and the coefficients tend toward zero (i.e., words are not informative of a document's structured data). In essence, the synthesizer varies its level of privacy protection with a single protection parameter λ, and it converges to random sampling as λ increases.

Importantly for computational efficiency, we note that the Bayesian posterior mode under independent Laplace (double exponential) distributions is equivalent to the LASSO estimate with the penalty term $\lambda \sum_{k_j=1}^{K_j} |\boldsymbol{u}_{k_j}|$ (Tibshirani (1996), p. 277). We use this relation for computational efficiency of the voluminous textual data (often in the thousands or tens of thousands of variables) in our real data application in Section 4. Our synthesis goal is to reduce the coefficients to zero when words are overly informative to each structured data covariate (and hence, a specific user in a small area) and compromising user privacy. To ensure that a high value of λ (e.g., $\lambda = 0.5$) results in an equal probability of choosing any level k_j, we do not include an intercept in the model. This implies that our data protection model has "privacy by design" in accordance with the European Union's General Data Protection Regulation[3].

[3] For more information, see `https://gdpr-info.eu/issues/privacy-by-design/`

Once the series of sequential Bayesian GLM synthesizers with shrinkage prior are estimated on the confidential data, we can use the estimated model parameters to simulate synthetic values of structured data levels. To do so for the level of document i covariate j, we sample from its posterior predictive density, conditional on its un-synthesized textual data $\boldsymbol{x}_i$ and its *synthesized* categorical covariates from their own steps, $(y_{i1}^{\lambda}, y_{i2}^{\lambda}, \cdots, y_{i(j-1)}^{\lambda})$, where superscript λ denotes synthetic values generated with privacy parameter λ in Equation (2). Again for notation simplicity, we use $\tilde{\boldsymbol{x}}_i^{\lambda}$ to denote the collection of the un-synthesized text data $\boldsymbol{x}_i$ and synthesized $(y_{i1}^{\lambda}, y_{i2}^{\lambda}, \cdots, y_{i(j-1)}^{\lambda})$. The posterior predictive density for y_{ij}^{λ} has the form:

$$p(y_{ij}^{\lambda} \mid \boldsymbol{D}, \tilde{\boldsymbol{x}}_i^{\lambda}, \lambda) = \int_{\boldsymbol{u}_{k_j}} p(y_{ij}^{\lambda} \mid \boldsymbol{u}_{k_j}, \tilde{\boldsymbol{x}}_i^{\lambda}) p(\boldsymbol{u}_{k_j} \mid \boldsymbol{D}, \lambda) d\boldsymbol{u}_{k_j}, \tag{3}$$

where $p(\boldsymbol{u}_{k_j} \mid \boldsymbol{D}, \lambda)$ is the posterior density of the textual coefficients and other structured data covariates given the confidential data $\boldsymbol{D} = [\boldsymbol{X}, \boldsymbol{Y}]$ and conditional on the privacy parameter λ. When $M > 1$ partially synthetic datasets are simulated, we label them $\boldsymbol{D}^{\lambda} = [\boldsymbol{D}^{\lambda,1}, \cdots, \boldsymbol{D}^{\lambda,M}]$, where the mth synthetic dataset is $\boldsymbol{D}^{\lambda,m} = [\boldsymbol{X}, \boldsymbol{Y}^{\lambda,m}] = [(\boldsymbol{x}_1, \boldsymbol{y}_1^{\lambda,m}), \cdots, (\boldsymbol{x}_n, \boldsymbol{y}_n^{\lambda,m})]$.

3 Utility and disclosure risks measures

3.1 Utility measures

Our utility measures consist of two types: (a) global utility measures, where we focus on evaluating the preservation of overall distributional characteristics of the synthetic data, and (b) analysis-specific utility measures, where we focus on comparing results of inferences conducted on the confidential data and those on the synthetic data.

3.1.1 Global utility measures

In our setup, the synthetic covariates only include structured data covariates, all of which categorical. Therefore, for global utility measures, we focus on comparing the cell-level relative difference of one-way and two-way tables, constructed based on the categorical structured data covariates in the confidential dataset and in *each* synthetic dataset.

For one-way table of categorical covariate j with K_j levels, we calculate the frequency of each level k_j of covariate j in the confidential data of $\boldsymbol{Y}$, as in $f_{j,k_j,\boldsymbol{Y}} = \sum I(y_{ij} = k_j)/n$, where $I(\cdot)$ is the indicator function and n is the sample size. Similarly, we calculate the frequency of each level k_j of covariate j in the mth synthetic data of $\boldsymbol{Y}^{\lambda,m}$, as in $f_{j,k_j,\boldsymbol{Y}^{\lambda,m}} = \sum I(y_{ij}^{\lambda,m} = k_j)/n$. We then report the relative difference $\delta_{j,k,m}$ specific to the mth synthetic data for covariate j and level k_j:

$$\delta_{j,k_j,m} = \frac{f_{j,k_j,\boldsymbol{Y}} - f_{j,k_j,\boldsymbol{Y}^{\lambda,m}}}{f_{j,k_j,\boldsymbol{Y}}}. \tag{4}$$

For example, for one-way table of synthetic state covariate of 8 levels, we calculate relative frequencies of each state level between the confidential and one synthetic datasets, and report 8 relative differences.

We define relative difference for two-way tables by taking cross tabulations, and we omit the expression for brevity. For example, for two-way table of synthetic state (8 levels) and synthetic rating (5 levels), we calculate relative frequencies of each of the $8 \times 5 = 40$ cells between the confidential and the synthetic datasets, and report 40 relative differences. When $M > 1$ synthetic datasets are generated, results from each synthetic dataset are reported. That is, if $M = 20$ synthetic datasets are generated, we report $M \times 8$ one-way relative differences for the synthetic state covariate, and $M \times 40$ two-way relative differences for the two-way tables of synthetic state and synthetic rating.

In general, the closer the relative difference is to 0, indicating smaller deviation of the cell count in the synthetic data from the confidential data, the higher the global utility is. Similar measures are used in Drechsler and Hu (forthcoming) and Hu and Savitsky (2018), which have been shown as good global utility measures for synthetic data containing categorical covariates.

3.1.2 Analysis-specific utility measures

To measure analysis-specific utility of textual data combined with structured data, we study how textual insights change when the structured data is protected. We use a Structural Topic Model (STM) from Roberts et al. (2014) to process high volumes of textual reviews heuristically using T topics, each of which is defined as a collection of words in a vocabulary $v = 1, \cdots, V$. The STM is an unsupervised model that discovers topics instead of assuming them *a priori*. It is also an extension of the Latent Dirichlet Allocation (LDA) model that analyzes how the prevalence and content of topics vary over structured covariates such as state and rating (Blei et al., 2003).

Prevelence coefficients enable a data user to discover topics that are more common to structured data levels. This is important because topics can have different meanings when they are associated with different levels such as low- or high-star ratings. For example, a topic that includes the words "price" and "pay" would have very different meanings for a product with a 1-star review versus a product with 5-star review. Prevalence coefficients are computed from a logistic normal generalized linear model and measure how often a level $k_j = 1, ..., K_j$ of covariate j mentions a topic $t = 1, \cdots, T$. If all J covariates are included as prevalence covariates, there are $(K_1 - 1) + (K_2 - 1) + \cdots + (K_J - 1)$ estimated regression coefficients plus the intercept for each topic. We define the estimated prevalence coefficients as $\mu_{k_j}^{(t)}$ for the confidential data and $\mu_{k_j}^{\lambda,(t),m}$ for the synthetic data with protection level λ. Once $M > 1$ partially synthetic datasets are generated, confidence intervals of coefficients can be calculated using combining rules (Reiter and Raghunathan, 2007; Drechsler, 2011), which is a widely-used analysis-specific utility measure.

Content coefficients are multinomial logit on words and measure how often words are used within topic t for level k_j of covariate j. We define $\boldsymbol{\beta}_{k_j} = [\boldsymbol{\beta}_{k_j}^{(1)}, \dots, \boldsymbol{\beta}_{k_j}^{(T)}]$ as the $V \times T$ matrix of content coefficients for level k_j of covariate j computed from the confidential data. Each entry in the matrix can be viewed as the importance of word v in topic t. The columns of $\boldsymbol{\beta}_{k_j}$ sum to one and contain the word-specific distribution for topic t. Define $\boldsymbol{\beta}_{k_j}^{\lambda,m} = [\boldsymbol{\beta}_{k_j}^{\lambda,\,(1),m}, \dots, \boldsymbol{\beta}_{k_j}^{\lambda,\,(T),m}]$ as the $V \times T$ matrix of content coefficients computed from the synthetic data with protection level λ.

To measure analysis-specific utility across all documents for topic t and level k_j of covariate j, we compute the correlation between $\boldsymbol{\beta}_{k_j}^{(t)}$ in the confidential data and $\boldsymbol{\beta}_{k_j}^{\lambda,\,(t),m}$ in the synthetic data,

$$Cor(\boldsymbol{\beta}_{k_j}^{(t)}, \boldsymbol{\beta}_{k_j}^{\lambda,(t),m}). \tag{5}$$

This results in $K \times T$ correlations for each synthetic dataset. A correlation coefficient close to 1 implies that textual documents with the same level of a covariate discuss topics in similar ways. A correlation value close to 0 indicates that the level-specific words in each topic degrade to “pure noise” in the English language. However, this situation is unlikely because textual documents are usually analyzed in topic models with the same product, service, or company. Thus, randomness in the English vocabulary is neither expected nor desired. The average correlation for topic t, covariate j, and protection level λ is defined as

$$\frac{1}{M \times K_j} \sum_{k_j=1}^{K_j} \sum_{m=1}^{M} Cor(\boldsymbol{\beta}_{k_j}^{(t)}, \boldsymbol{\beta}_{k_j}^{\lambda,(t),m}), \tag{6}$$

which is another analysis-specific utility measure we report.

3.2 Disclosure risks measures

As before, we label $\boldsymbol{D} = [\boldsymbol{X}, \boldsymbol{Y}] = [(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)]$ as the platform's confidential dataset with each row $\boldsymbol{d_i} = (\boldsymbol{x}_i, \boldsymbol{y}_i)$ containing structured data and textual content belonging to user i. Previously, we labeled $\boldsymbol{D}^{\lambda,m} = [\boldsymbol{X}, \boldsymbol{Y}^{\lambda,m}] = [(\boldsymbol{x}_1, \boldsymbol{y}_1^{\lambda,m}), \cdots, (\boldsymbol{x}_n, \boldsymbol{y}_n^{\lambda,m})]$ as the mth partially synthetic and publicly released dataset, with each row $(\boldsymbol{x}_i, \boldsymbol{y}_i^{\lambda,m})$ containing the un-synthesized textual data $\boldsymbol{x}_i$ and synthesized structured data $\boldsymbol{y}_i^{\lambda,m}$ of document i. For notation simplicity, we dropped the superscript m in $\boldsymbol{D}^{\lambda,m}$ and $\boldsymbol{Y}^{\lambda,m}$ and work with $\boldsymbol{D}^{\lambda}$ and $\boldsymbol{Y}^{\lambda}$ as this risk evaluation process is done on each of the $m = 1, \cdots, M$ synthetic datasets when $M > 1$.

We focus on identification disclosure risks measures and evaluation, where the intruder attempts to identify a record of interest with access to additional information about this record through external databases. Our approach follows and extends that of Drechsler and Reiter (2008); Reiter and Mitra (2009); Hu (2019). We assume that a data intruder wishes to identify a data record i in the publicly available and synthetic $\boldsymbol{D}^{\lambda}$, with one piece of external data that includes structured data with a writing sample (e.g., from emails, voice transcriptions, textual reviews on different platforms, data purchased from a data broker, or scraped data on other platforms) corresponding to an identified person. We label the intruder's external data as $\boldsymbol{d}_z = (\boldsymbol{x}_z, \boldsymbol{y}_z)$ where $\boldsymbol{x}_z$ and $\boldsymbol{y}_z$ are the textual data and structured data belonging to a target z. Let ID_z be the identity of target user z and ID_i be the identity of the user for document i. The task of the data intruder is to estimate the probability of $ID_i = ID_z$ given the synthetic data $\boldsymbol{D}^\lambda$ and target $\boldsymbol{d}_z = (\boldsymbol{x}_z, \boldsymbol{y}_z)$, namely,

$$Pr(ID_i = ID_z \mid \boldsymbol{D^\lambda}, \boldsymbol{d}_z) = \int_{(\boldsymbol{x}_i, \boldsymbol{y}_i^\lambda)} Pr \quad (ID_i = ID_z \mid \boldsymbol{x}_i, \boldsymbol{y}_i^\lambda, \boldsymbol{D^\lambda}, \boldsymbol{d}_z)$$
$$Pr \quad (\boldsymbol{x}_i, \boldsymbol{y}_i^\lambda \mid \boldsymbol{d}_z, \boldsymbol{D^\lambda}) d(\boldsymbol{x}_i, \boldsymbol{y}_i^\lambda), \tag{7}$$

where $(\boldsymbol{x}_i, \boldsymbol{y}_i^\lambda)$ is the synthetic data record i in $\boldsymbol{D}^\lambda$. The data intruder repeats this probability estimation process for each of the n records in $\boldsymbol{D}^\lambda$.

Recall that in our approach, textual data $\boldsymbol{x}_i$ is un-synthesized while the structured data $\boldsymbol{y}_i^\lambda$ is synthesized. Therefore, we propose a two-step approach to estimating the probability of identification in Equation (7): first matching with structured data $\boldsymbol{x}_z$, and next matching with textual data $\boldsymbol{y}_z$.

In the first stage, the data intruder retains potential users $i \in (1, \ldots, n)$ in synthetic $\boldsymbol{D}^\lambda$ whose structured data $\boldsymbol{y}_i^\lambda$ match $\boldsymbol{y}_z$, the structured data of target z from external data. For example, if the data intruder knows the state of target z is New Mexico, she may retain all users in $\boldsymbol{D}^\lambda$ whose $\boldsymbol{y}_i^\lambda$ includes New Mexico as their state. If in addition, the data intruder knows the rating of target z is 3, then she may retain all users in $\boldsymbol{D}^\lambda$ whose $\boldsymbol{y}_i^\lambda$ are New Mexico *and* rating of 3. Based on matching with structured data $\boldsymbol{y}_z$ with $\boldsymbol{y}_i^\lambda$ in $\boldsymbol{D}^\lambda$, the intruder assigns a first-stage probability of identification,

$$Pr(ID_i = ID_z \mid \boldsymbol{D^\lambda}, \boldsymbol{y}_z) = \frac{1}{N_z}, \tag{8}$$

where $N_z^\lambda = \{i \mid \boldsymbol{y}_i^\lambda = \boldsymbol{y}_z\}$ is the collection of records with $\boldsymbol{y}_i^\lambda = \boldsymbol{y}_z$ and $N_z = |N_z^\lambda|$, its carnality.

In the second stage, the data intruder computes the second-stage probability of identification by matching the textual data $\boldsymbol{x}_i$ of documents $i \in N_z^\lambda$ to $\boldsymbol{x}_z$. We define a distance-based metric

$$d(\boldsymbol{x}_i, \boldsymbol{x}_z) = \left|\left| \frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i}{\sigma_{\boldsymbol{x}_i}} - \frac{\boldsymbol{x}_z - \bar{\boldsymbol{x}}_z}{\sigma_{\boldsymbol{x}_z}} \right|\right|_1, \tag{9}$$

to be the l_1-norm distance between the standardized textual data (where the frequency of each word in document i is standardized). Define $s(\boldsymbol{x}_i, \boldsymbol{x}_z) = 1/d(\boldsymbol{x}_i, \boldsymbol{x}_z)$ to be the similarity and the data intruder normalizes all N_z^λ similarities to sum to unity. The second-stage probability of identification is then

$$Pr(ID_i = ID_z \mid \boldsymbol{D^\lambda}, \boldsymbol{y}_z, \boldsymbol{x}_z) = \begin{cases} \frac{s(\boldsymbol{x}_i, \boldsymbol{x}_z)}{\sum_{i' \in N_z^\lambda} s(\boldsymbol{x}_{i'}, \boldsymbol{x}_z)} & \text{if } \ i \in N_z^\lambda \\ 0 & \text{if } \ i \notin N_z^\lambda \end{cases}, \tag{10}$$

which is the final identification probability for data record i we report as our disclosure risks measures.

Our presentation and description of the identification risk evaluation focuses on the case where $M = 1$, i.e. one synthetic dataset is simulated. However, the platform typically simulates $M > 1$ synthetic datasets to compute the probability of identification and choose the optimal protection level, λ. This is important so that the results are not dependent on a single lucky (or unlucky) draw of synthetic data. For example, there would be a very different probability of identification if $\boldsymbol{y}_i$ was changed only once from Georgia to Texas, instead of $M > 1$ times to all states including Georgia. When $M > 1$, we estimate $Pr(ID_i^m = ID_z \mid \boldsymbol{D}^{\lambda,m}, \boldsymbol{d}_z)$ for $m = 1, \cdots, M$ and take the average identification probability across $M > 1$ calculated probabilities, as our final and reported identification probability for data record i. In practice, the platform can run an internal privacy assessment and compute this probability on their own by choosing a large enough M.

4 Application to Consumer Affairs reviews

4.1 The data

We use data from ConsumerAffairs.com to demonstrate our methodology in the substantive area of online reviews. ConsumerAffairs.com is a popular online review platform with over one million reviews and 3,000 brands. We collected 1,921 reviews from 1,881 unique reviewers (users) on August 21, 2017 from one financial services vendor in the United States that offers insurance products. Near the date of data collection, ConsumerAffairs.com positioned itself as a negative reviews platform where companies could pay to "convert detractors into brand promoters."[4] Our financial services vendor had an average star rating of 2.1 stars which was similar to many other popular companies. However, in 2021, ConsumerAffairs.com appears to have rebranded itself resulting in only 1463 reviews for our vendor, an average star rating closer to 4 stars, and clear disclosures to consumers on how ConsumerAffairs.com makes money.

In order to submit a review in our collected dataset, the platform required a reviewer to submit their email address, phone number, full mailing address, first name, last name, and an optional interest group(s) (e.g., homeowner, pet parent, silver senior, etc.). However, the platform did not display all this information publicly (or to non-paying vendors) – the published structured data of each review included the reviewer's first name, date of review, city, state, and star rating. Additionally, the reviewer could choose to alter their displayed first name to their first initial by changing their public alias on their profile.

The textual content of each review contains between 200 and 50,000 characters in free text form. The textual variables used in our application include the frequency of any word that appears (thousands of variables), the frequency of ! . ? # @ $

[4] https://www.consumeraffairs.com/brands/convert-detractors-2 (accessed January 7th, 2018)

% & , ; : ( ) (13 variables), the frequency of words like “the”, “of”, “and” (174 variables), the frequency of a to z (ignoring case) (26 variables), the frequency of 0 to 9 (10 variables), the number of words (1 variable), the frequency of words with a mix of upper and lower case letters (1 variable), the frequency of words that have 1-30 characters (30 variables), the vocabulary richness (Yule’s K; Yule (2014)) (1 variable), and the Flesch-Kinkaid reading level (Kincaid et al., 1975) (1 variable). Note that only the frequency of any word that appears is used to measure the data utility, but all variables are used to measure the disclosure risks of our simulated synthetic data.

For our application, we define each review i having structured data $\boldsymbol{y}_i$ consisting of the two covariates ($J = 2$) of star rating (y_{i1}) and state (y_{i2}), and textual content $\boldsymbol{x}_i$. We use $\boldsymbol{X}$ and $\boldsymbol{Y}$ to denote the textual data and structured data for the confidential dataset $\boldsymbol{D} = [\boldsymbol{X}, \boldsymbol{Y}]$.[5] We use $K_1 = 5$ levels for star rating (1, 2, 3, 4, 5) and reduce state to $K_2 = 8$ levels (New Mexico, Georgia, Florida, California, North Carolina, Texas, Virginia, and Illinois), resulting in $n = 821$ reviews for $\boldsymbol{D}$. Separately, we withhold 9 target reviews $\boldsymbol{d}_z = (\boldsymbol{x}_z, \boldsymbol{y}_z)$ with structured data and textual content for targets $z = 1, \ldots, 9$. These target reviews are secondary reviews *not* in $\boldsymbol{D}$ but from 9 of the same users in $\boldsymbol{D}$.

4.2 The synthesis

We consider two types of synthesis for the 821 reviews in the confidential dataset $\boldsymbol{D}$. The first type synthesizes one categorical covariate, the state, while leaving rating unchanged. This synthesis type is labeled as “State Only”. The second synthesis type not only synthesizes state but also rating. That is, we perform a sequential synthesis of synthesizing state first and rating second. This synthesis type is labeled

[5] We acknowledge that synthesizing star rating is likely infeasible for the platform due to user experience concerns, but we synthesize the rating to demonstrate that our methodology can accommodate multiple structured covariates when present in the data.

as “State and Rating.”

For the “State Only” synthesis, we fit confidential state y_{i2} with $K_2 = 8$ levels of $\{\text{New Mexico}, \ldots, \text{Illinois}\}$ as the dependent variable using our proposed Bayesian GLM synthesizer in Section 2, conditional on confidential rating y_{i1} and textual content $\boldsymbol{x}_i$ consisting of 6,767 textual variables. Since $\boldsymbol{x}_i$ is sparse and the number of textual variables far exceeds the number of reviews, we use the LASSO estimate λ for computational efficiency. We experiment different values of λ, $\lambda = \{10^{-4}, 0.03, 0.5\}$, to evaluate its effect on the utility and disclosure risks evaluations of the resulting synthetic data. According to our proposed Bayesian GLM synthesizer and the LASSO estimate setup discussed in Section 2, smaller λ provides less smoothing, which should result in synthetic data with higher utility and higher disclosure risks.

After fitting the Bayesian synthesis model on the confidential dataset, we simulate a synthetic value of y_{i2}^{λ} of review i, $i = 1, \cdots, n$, resulting in one partially synthetic dataset. We repeat this process for $M = 20$ times to obtain $M = 20$ synthetic datasets. Each protected “State Only” dataset $\boldsymbol{D}^{\lambda,m}$ consists of unaltered textual content $\boldsymbol{x}_i$, synthesized state y_{i2}^{λ}, and confidential rating y_{i1} of review i and for all $n = 821$ reviews.

For the “State and Rating” sequential synthesis, we first fit confidential rating y_{i1} with $K_1 = 5$ levels of $\{1, \ldots, 5\}$ as the dependent variable using our proposed Bayesian GLM synthesizer on the textual content $\boldsymbol{x}_i$ only. We generate a synthetic value of y_{i1}^{λ}, for all $i = 1, \cdots, n = 821$ reviews. As with “State Only” synthesis, we use $\lambda = \{10^{-4}, 0.03, 0.5\}$. We then fit a GLM with a dependent variable of confidential state y_{i2} conditional on the confidential rating y_{i1} and textual content $\boldsymbol{x}_i$ with $\lambda = \{10^{-4}, 0.03, 0.5\}$. We then simulate a synthetic value of y_{i2}^{λ}, for all $i = 1, \cdots, n = 821$ reviews. As before, we repeat the process for $M = 20$ times to obtain $M = 20$ synthetic datasets. Each protected “State and Rating” synthetic dataset $\boldsymbol{D}^{\lambda,m}$ consists of unaltered textual content $\boldsymbol{x}_i$, synthesized rating y^{i1}, and synthesized state y_{i2}^{λ} of review i and for all $n = 821$ reviews.

4.3 Utility results

4.3.1 Global utility results

We report the global utility introduced in Section 3.1.1, where we calculate the relative differences of one-way and two-way frequencies in the confidential dataset and the M synthetic datasets of each synthesis type. Figure 1 presents a density plot of the relative differences of $M = 20$ synthetic datasets and Table 1 reports the average absolute difference over $M = 20$ datasets, for “State Only” synthesis and “State and Rating” sequential synthesis.

The results show that as λ decreases, the one-way and two-way relative differences decrease, regardless of the synthesis type. These indicate that our privacy parameter, λ, helps tune the level of utility preservation, in the intended direction: smaller λ induces less smoothing, resulting in synthetic data with higher utility (i.e. smaller one-way and two-way relative differences). These results also show minimal differences between “State Only” synthesis and “State and Rating” sequential synthesis for the one-way tables. However, the global utility of the two-way tables for “State and Rating” sequential synthesis are about twice as large as the “State Only” synthesis. Since the larger the average absolute difference the lower the utility, “State and Rating” sequential synthesis provides lower utility, which is expected since synthetic state values are simulated based on synthetic rating values in the sequential synthesis approach.

Moreover, for a high level of protection with $\lambda = 0.5$ with “State and Rating” sequential synthesis, the bottom right panel of Figure 1 shows a negative bias, indi-

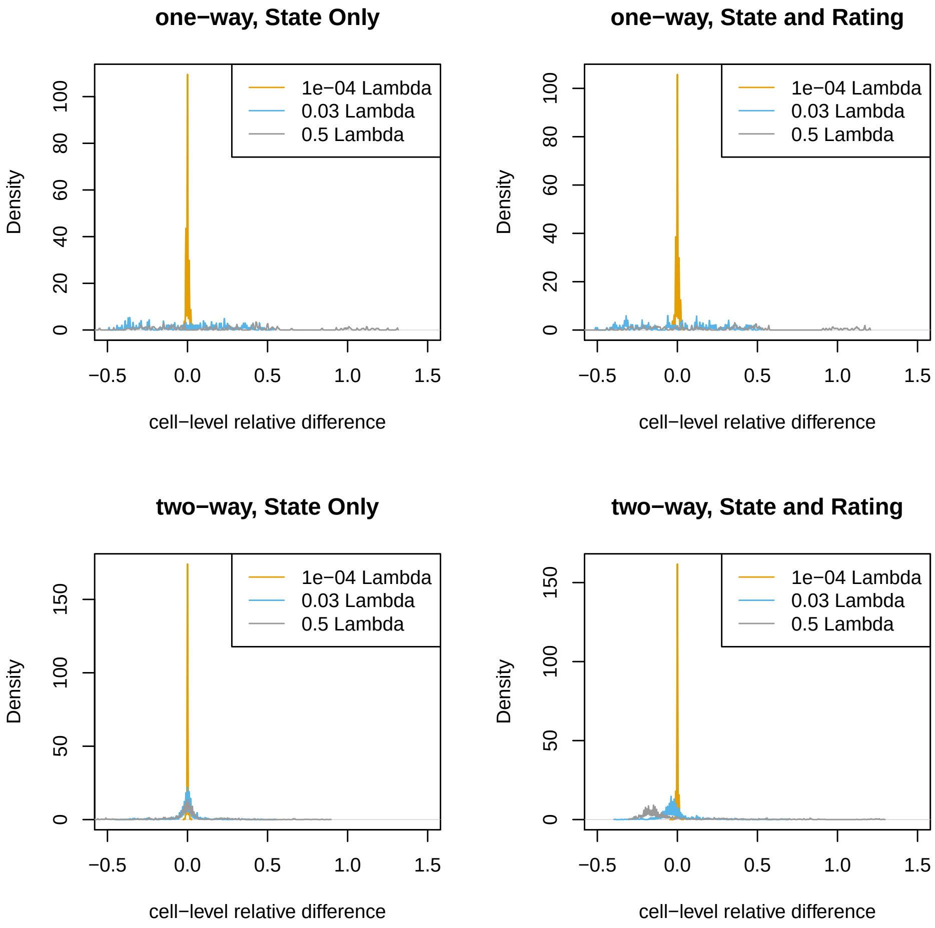


Figure 1: Density plots of relative differences of one-way and two-way frequencies as global utility for $M = 20$ synthetic datasets, of “State Only” and “State and Rating”.

cating that most of the two-way counts in the confidential data are larger than the counts in the synthetic datasets. This makes sense because as the privacy level λ increases, there is a more similar probability of choosing each state and rating. Most of the counts in the synthetic dataset increase because the confidential dataset has only a few large counts with the majority being below average (lower probability of a review occurring in a specific state with a specific rating).

	State Only		State and Rating	
	One-way	Two-way	One-way	Two-way
$\lambda = 0.5$	3.99	3.93	4.34	9.09
$\lambda = 0.03$	1.86	1.84	2.42	3.62
$\lambda = 1e0 - 4$	0.04	0.05	0.05	0.08

Table 1: Average absolute differences as global utility for $M = 20$ synthetic datasets, of "State Only" and "State and Rating".

Topic	Topic Words	High Rating Prevalence in CA	Low Rating Prevalence in CA	Low Rating Coefficient for all states
1	processors, med, processor	0.039	0.097	0.058
2	ticketed, police, driver	0.170	0.280	0.110
3	discounts, easy, reasonable	0.650	0.140	-0.510
4	debt, funds, credit	0.089	0.326	0.237
5	sat, truck, shop	0.051	0.156	0.105
Total		1.000	1.000	0.000

Table 2: Prevalence in California of high rating (4 or 5 stars) and low rating (1 to 3 stars), and low rating coefficient for all states in the confidential data.

4.3.2 Analysis-specific utility results

For analysis-specific utility, we select $T = 5$ topics and discretize rating into low rating (1 to 3 stars) and high rating (4 or 5 stars). Table 2 shows the top 3 topic words across all states, the high or low rating prevalence in California, and the low rating coefficient for all states for the *confidential* data. Topic 3 is predominantly a high rating topic for all topics (e.g., the low rating coefficient is -0.510 and represents a deviation in the prevalence between high rating and low rating) and is represented by the collection of the words "discounts, easy, and reasonable." The other four topics have positive low rating coefficients across all states and consist of mostly criticisms of the insurance product. For example, Topic 2 is composed of the words "ticketed, police, and driver" which likely describes auto insurance problems resulting from a traffic ticket.

	point estimate	95% C.I.
Conf. Data	0.110	[0.056, 0.163]
$\lambda = 0.5$	-0.002	[-0.053, 0.050]
$\lambda = 0.03$	0.017	[-0.054, 0.088]
$\lambda = 1e0 - 4$	0.114	[0.060, 0.168]

Table 3: Table of C.I.s of Topic 2 estimateof the low rating coefficient for "State and Rating" synthesis with $M = 20$ synthetic datasets.

	point estimate	95% C.I.
Conf. Data	-0.510	[-0.563, -0.456]
$\lambda = 0.5$	0.000	[-0.047, 0.048]
$\lambda = 0.03$	-0.180	[-0.263, -0.098]
$\lambda = 1e0 - 4$	-0.507	[-0.562, -0.452]

Table 4: Table of C.I.s of Topic 3 estimate of the low rating coefficient for "State and Rating" synthesis with $M = 20$ synthetic datasets.

As one of our analysis-specific utility measures, we present the change in the low rating coefficient for the "State and Rating" synthesis. Tables 3 and 4 show the low rating coefficient for Topic 2 and Topic 3. For low privacy levels with $\lambda = 1e0-4$ (i.e., 0.0001), both topics have point estimates similar to those in the confidential data. However, the low rating coefficient to shrink toward 0 for a medium privacy level of $\lambda = 0.03$ and high privacy level of $\lambda = 0.5$. The reasoning is that higher privacy levels cause the synthesized rating to approach pure noise. Results are similar when analyzing confidence intervals across other topics. Additionally, we analyzed the confidence intervals for the "State Only" synthesis and found that the confidential point estimate of the Topic 2 low rating coefficient in Table 3 are within all the 95% confidence intervals. The confidential point estimate of Topic 3 low rating coefficient is within the 95% confidence interval of $\lambda = 1e0 - 4$ and $\lambda = 0.03$ for "State Only" synthesis.

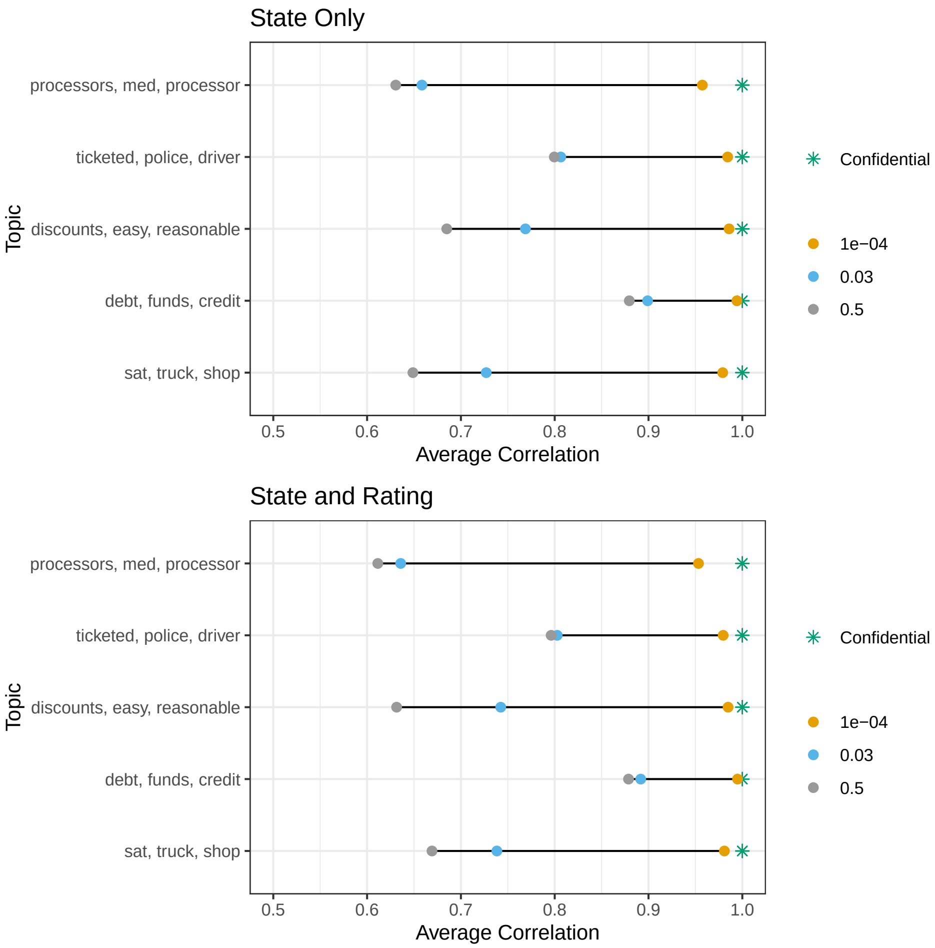
Our other analysis-specific utility measure is content coefficients, which measure how often all words (not only the top 3) are used in topic t for specific states. Figure 2 illustrates the average correlation of the content coefficients within states between the confidential dataset and the synthesized datasets. The average is calculated across all $K_2 = 8$ states and $M = 20$ synthetic datasets. The results show that low privacy levels with $\lambda = 1e0 - 4$ barely change how often words are used in topic t for specific levels of state. For example, the words "discounts, easy, and reasonable" in Topic 3 are used with similar frequencies in both the confidential data and the synthetic data with $\lambda = 10^{-4}$, i.e. $\{\boldsymbol{D}^{1e0-4,1}, \ldots, \boldsymbol{D}^{1e0-4,20}\}$. It also appears that the average correlations are usually lower for "State and Rating" synthesis compared to "State Only" synthesis because of the extra synthetic data step in the process.

For medium to high privacy levels, the average correlations drop to between 0.60 and 0.90 depending on the topic. We note that the underlying correlations do not drop to 0 or a negative value because only the structured data are synthesized. The textual content are not synthesized and all relate to the same insurance products. However, we could expect the average correlation to drop to 0 if we replaced textual content in each review with an incoherent random sample of the 6,767 textual variables.

In summary, our global utility and analysis-specific utility evaluations of "State Only" and "State and Rating" syntheses demonstrate that our single privacy parameter, λ, in our proposed Bayesian GLM synthesizer, can fine tune the utility preservation level of the resulting synthetic data. To reach high level of utility preservation, we can choose a small λ. We now turn to our disclosure risks evaluation results.

4.4 Disclosure risks results

As presented in Section 3.2, for disclosure risks evaluation, we report the average second-stage probabilities of identification using both the structured data and textual content for the true match to target z. The green stars in Figure 3 show the


State Only

State and Rating

Figure 2: Average correlations of content coefficients, between confidential dataset and $M = 20$ synthetic datasets, of “State Only” synthesis and “State and Rating” synthesis.

disclosure risks inherent in the confidential data: Each green star is the probability of identification for one of the 9 target reviews when matching it to the confidential data $\boldsymbol{D}$ based on state, rating, and textual content. The textual content is different between the target review and the matched review, but they are written by the same author. The combination of matching with the structured data and textual content results in sizeable probabilities of identification from a population of 821 users. For example, the identification probability of “ZITA” in the confidential data $\boldsymbol{D}$ is close to 0.1.

We evaluate the identification disclosure probabilities of these 9 target records in synthetic datasets, generated using our Bayesian GLM synthesizer with λ privacy parameter. For the synthetic datasets $\{\boldsymbol{D}^{\lambda,1}, \ldots, \boldsymbol{D}^{\lambda,M}\}$ ($M = 20$), as λ increases, the identification probability of each target decreases, as shown in Figure 3. For most target records, synthetic data simulated with $\lambda = 1e0 - 4$ has identification probabilities as high as the confidential data. The identification probabilities substantially decrease as λ increases, and for most cases, the difference between $\lambda = 0.03$ and $\lambda = 0.5$ is small.

These results demonstrate that our proposed Bayesian GLM synthesizer with single privacy parameter λ fine tune the disclosure risks of the resulting synthetic data. As λ increases with less smoothing induced in the synthetic data, the number of potential matches N_z^λ increases, which leads to a lower first-stage probability of identification, $1/N_z^\lambda$ in Equation (8) for target z. This results in more matches for the target's textual content $\boldsymbol{x}_z$ to textual content in $\boldsymbol{X}$, which decreases the second-stage probability of identification, our reported identification probability. Furthermore, we find that as N_z^λ reaches about 100 or more with $\lambda \geq 0.03$, the added gain by the second-stage is negligible due to the abundance of textual comparisons, resulting in similar performance between $\lambda = 0.03$ and $\lambda = 0.5$.

We also conducted a sensitivity analysis of M, the number of synthetic datasets, where $M = \{2, 3, 5, 20, 50\}$. We have found that when M is sufficiently large, e.g. $M = 20$ in our application, the utility results and disclosure risks results stabilized. We include results of $M = 50$ in the Appendix for comparison. We advocate for sufficiently large M, but not too large to avoid unnecessary computational burden.

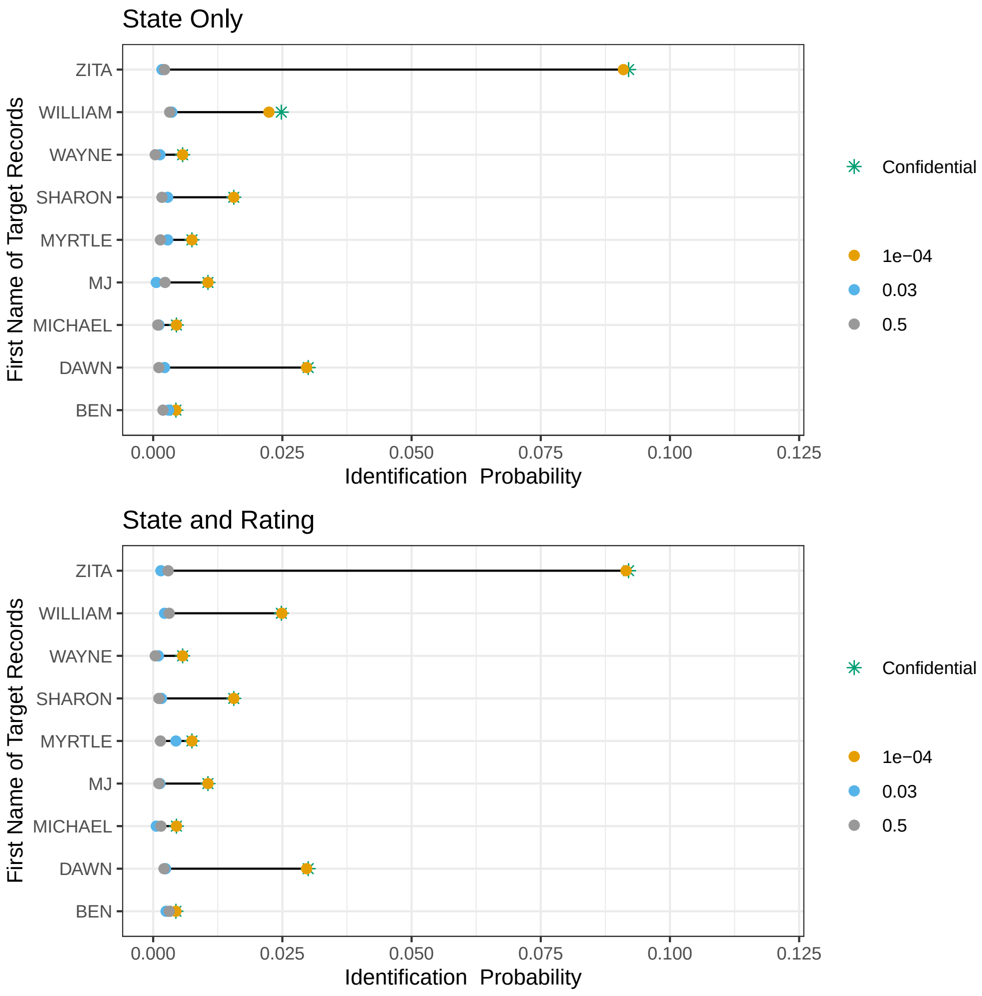


Figure 3: Identification probability for each of the 9 target record in the $M = 20$ synthetic datasets, of "State Only" synthesis and "State and Rating" synthesis.

5 Concluding remarks

Our proposed data protection framework through Bayesian data synthesis applies to UGC platforms publishing users' structured data alongside their textual content. As platforms have put the onus on users to protect themselves, the potential for such low level of privacy protection implies an evolving privacy landscape in the UGC industries, and thus strongly motivated our investigation.

The focus of our study was how a platform can implement a data protection policy for structured data and textual content at the user level. We found that our proposed Bayesian data synthesis model that explicitly conditions on the textual content to generate the synthetic structured data was effective at (a) reducing the probability of identifying an anonymous user and (b) maintaining the distributional characteristics and the topical content in structured data. To control for privacy, we used a Laplace prior distribution with privacy protection parameter λ that scaled back the amount of textual content contained in the posterior predictive density of the structured data. At moderate levels of protection, e.g. $\lambda = 0.03$ for the Consumer Affairs data sample, the average probability of identification for an anonymous decreased substantially while maintaining most of the distributional characteristics and topical insights in the structured data. Such a data protection policy can increase data privacy protection, effectively reducing the expected benefit of taking action against potential users so much that it would likely stop the behavior entirely. Furthermore, our proposed Bayesian data synthesizer with a shrinkage prior mathematically employed "privacy by design" which is an upcoming mandate in data protection legislation.

There are notable limitations in our paper. First, note that a user can unwittingly significantly increase their chances of being identified by providing extra information within textual content that links to their structured data. For example, they may say "Yesterday, I called my insurance company about an accident" which would give the data intruder an exact date of their phone call. This extra piece of structured data contained within the text extends beyond the scope of our Bayesian data synthesizer and is an interesting future direction. Second, our study likely understates the lack of data privacy on other UGC platforms with more textual content because each targeted user in our application had only two pieces of textual content. Prior research shows that data privacy protection decreases dramatically when the textual content on a user contains more training data (user reviews on IMDb dataset, see Qian et al. (2014)) or over 6,500 words (student essays, see Brennan et al. (2012)).

We note that our proposed data protection framework through Bayesian data synthesis is not limited to the context of online reviews. For example, regulation in the pharmaceutical industry around protection of Clinical Study Reports, which contain both structured data and textual data on clinical trial participants and staff, suggests maximum probability of re-identification of 9% for protected data (European Medicines Agency, 2018). By using our methods to incorporate a synthesizer for the combination of textual and structured data, pharmaceutical companies can run an internal privacy assessment and estimate this probability by repeating the synthesis process multiple times. Other examples where anonymity is important and our framework can be applied include federal user complaint databases, employee satisfaction surveys, teaching evaluations, peer reviews for academic journals, and online drug and alcohol support groups. While protecting the text itself is not appropriate in our context, this may be useful for other application areas. Automated text protection is a challenging problem (Potthast et al., 2018) that could have major privacy implications if solved. We encourage these areas for future methodology research and applications.

References

Beigi, G. and Liu, H. (2020) A survey on privacy in social media: Identification, mitigation, and applications. *ACM Transactions on Data Science*, **1**, 1–38.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Brennan, M., Afroz, S. and Greenstadt, R. (2012) Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, **15**, 1–22.

Burgette, L. F. and Reiter, J. P. (2013) Multiple-shrinkage multinomial probit models with applications to simulating geographies in public use data. *Bayesian Analysis*, **8**, 453–478.

Campbell, J., Goldfarb, A. and Tucker, C. (2015) Privacy regulation and market structure. *Journal of Economics & Management Strategy*, **24**, 47–73.

Drechsler, J. (2011) *Synthetic Datasets for Statistical Disclosure Control.* Springer: New York.

Drechsler, J. and Hu, J. (forthcoming) Synthesizing geocodes to facilitate access to detailed geographical information in large scale administrative data. *Journal of Survey Statistics and Methodology.*

Drechsler, J. and Reiter, J. P. (2008) Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases*, 227–238. Springer, Berlin, Heidelberg.

Duncan, G. T. and Lambert, D. (1986) Disclosure-limited data dissemination. *Journal of the American Statistical Association*, **10**, 10–28.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, 265–284. Berlin, Heidelberg: Springer-Verlag.

European Medicines Agency (2018) External guidance on the implementation of the european medicines agency policy on the publication of

clinical data for medicinal products for human use. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clini
en-3.pdf.

Feyisetan, O., Diethe, T. and Drake, T. (2019) Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, 210–219. IEEE.

Goldfarb, A. and Tucker, C. (2013) Why managing consumer privacy can be an opportunity. *MIT Sloan Management Review*, **54**, 10.

Hu, J. (2019) Bayesian estimation of attribute and identification disclosure risks in synthetic data. *Transactions on Data Privacy*, **12**, 61–89.

Hu, J. and Savitsky, T. D. (2018) Bayesian data synthesis and disclosure risk quantification: an application to the Consumer Expenditure Surveys. arXiv:1809.10074.

Jiménez, F. R. and Mendoza, N. A. (2013) Too popular to ignore: The influence of online reviews on purchase intentions of search and experience products. *Journal of Interactive Marketing*, **27**, 226–235.

Jones, R. (2017) Court rules yelp must identify anonymous user in defamation case. *Gizmodo*. URL: https://gizmodo.com/court-rules-yelp-must-identify-anonymous-user-in-defama-1820433103.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. and Chissom, B. S. (1975) Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Tech. rep.*, Naval Technical Training Command Millington TN Research Branch.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S. and Abowd, J. M. (2011) Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review*, **79**, 362–384.

Krishna, S., Gupta, R. and Dupuy, C. (2021) Adept: Auto-encoder based differentially private text transformation. *arXiv preprint arXiv:2102.01502*.

Kumar, V. and Reinartz, W. (2018) Customer privacy concerns and privacy protective responses. In *Customer Relationship Management*, 285–309. Springer.

Laskowski, N. (2014) Nordstrom digs into 5-star customer reviews and finds a shipping problem. URL: `https://searchcio.techtarget.com/feature/Nordstrom-digs-into-5-star-customer-reviews-and-finds-a-shipping-problem`.

Little, R. J. A. (1993) Statistical analysis of masked data. *Journal of Official Statistics*, **9**, 407–426.

Mankad, S., Han, H. Goh, J. and Gavirneni, S. (2016) Understanding online hotel reviews through automated text analysis. *Service Science*, **8**, 124–138.

Martin, K. D. and Murphy, P. E. (2017) The role of data privacy in marketing. *Journal of the Academy of Marketing Science*, **45**, 135–155.

Paiva, T., Chakraborty, A., Reiter, J. and Gelfand, A. (2014) Imputation of confidential data sets with spatial locations using disease mapping models. *Stat. Med.*, **33**, 1928–1945.

Porter, J. (2019) Fraudulent yelp posting protected under the law, ridiculous. *Tahoe Daily Tribune*. URL: `https://www.tahoedailytribune.com/news/jim-porter-fraudulent-yelp-posting-protected-under-the-law-ridiculous/`.

Potthast, M., Schremmer, F., Hagen, M. and Stein, B. (2018) Overview of the author obfuscation task at pan 2018: A new approach to measuring safety. In *CLEF (Working Notes)*.

Qian, T., Liu, B., Chen, L. and Peng, Z. (2014) Tri-training for authorship attribution with limited training data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 345–351.

Quick, H., Holan, S. H., Wikle, C. K., and Reiter, J. P. (2015) Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics*, **14**, 439–451.

Quick, H., Holan, S. H. and Wikle, C. K. (2018) Generating partially synthetic geocoded public use data with decreased disclosure risk by using differential smoothing. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **181**, 649–661.

Reiter, J. P. and Mitra, R. (2009) Estimating risks of identification disclosure in partially synthetic data. *The Journal of Privacy and Confidentiality*, **1**, 99–110.

Reiter, J. P. and Raghunathan, T. E. (2007) The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, **102**, 1462–1471.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., LederLuis, J., Gadarian, S. K., Albertson, B. and Rand, D. G. (2014) Structural topic models for open-ended survey responses. *American Journal of Political Science*, **58**, 1064–1082.

Schneider, M. J. and Abowd, J. M. (2015) A new method for protecting interrelated time series with bayesian prior distributions and synthetic data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 963–975.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodology)*, **58**, 267–288.

Wang, H. and Reiter, J. (2012) Multiple imputation for sharing precise geographies in public use data. *Ann. Appl. Stat.*, **6**, 229–252.

Wieringa, J., Kannan, P., Ma, X., Reutterer, T., Risselada, H. and Skiera, B. (2021) Data analytics in a privacy-concerned world. *Journal of Business Research*, **122**, 915–925.

Yule, C. U. (2014) *The statistical study of literary vocabulary.* Cambridge University Press.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D. and Xiao, X. (2017) Privbayes: Private data release via bayesian networks. In *ACM Transactions on Database Systems.*

Appendix

We include results with $M = 50$ for comparison with $M = 20$ in our Consumer Affairs application in Section 4. We report that all results for for $M = 50$ are qualitatively and quantitatively similar to the results for $M = 20$.

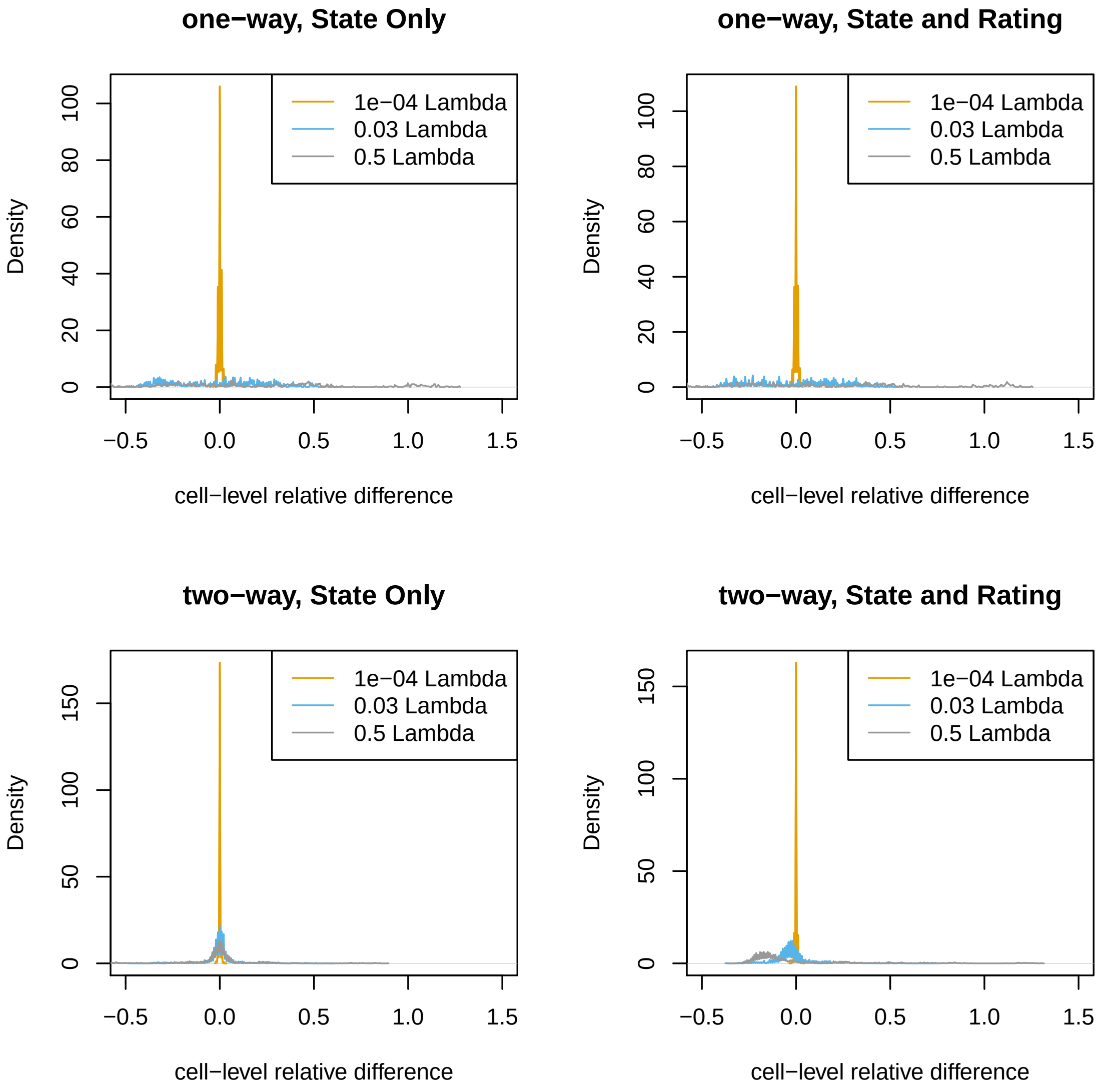


Figure 4: Density plots of relative differences of one-way and two0way frequencies as global utility for $M = 50$ synthetic datasets, of "State Only" and "State and Rating".

	State Only		State and Rating	
	One-way	Two-way	One-way	Two-way
$\lambda = 0.5$	4.00	4.02	4.36	9.06
$\lambda = 0.03$	1.80	1.69	2.37	3.47
$\lambda = 1e0 - 4$	0.04	0.04	0.06	0.08

Table 5: Average absolute differences for $M = 50$ synthetic datasets, of "State Only" and "State and Rating".

	point estimate	95% C.I.
Conf. Data	0.110	[0.056, 0.163]
$\lambda = 0.5$	-0.009	[-0.061, 0.043]
$\lambda = 0.03$	0.011	[-0.050, 0.073]
$\lambda = 1e0 - 4$	0.113	[0.059, 0.166]

Table 6: Table of C.I.s of Topic 2 estimate of the low rating coefficient for "State and Rating" synthesis with $M = 50$ synthetic datasets.

	point estimate	95% C.I.
Conf. Data	-0.510	[-0.563, -0.456]
$\lambda = 0.5$	0.007	[-0.036, 0.050]
$\lambda = 0.03$	-0.172	[-0.237, -0.107]
$\lambda = 1e0 - 4$	-0.507	[-0.561, -0.454]

Table 7: Table of C.I.s of Topic 3 estimate of the low rating coefficient for "State and Rating" synthesis with $M = 50$ synthetic datasets.

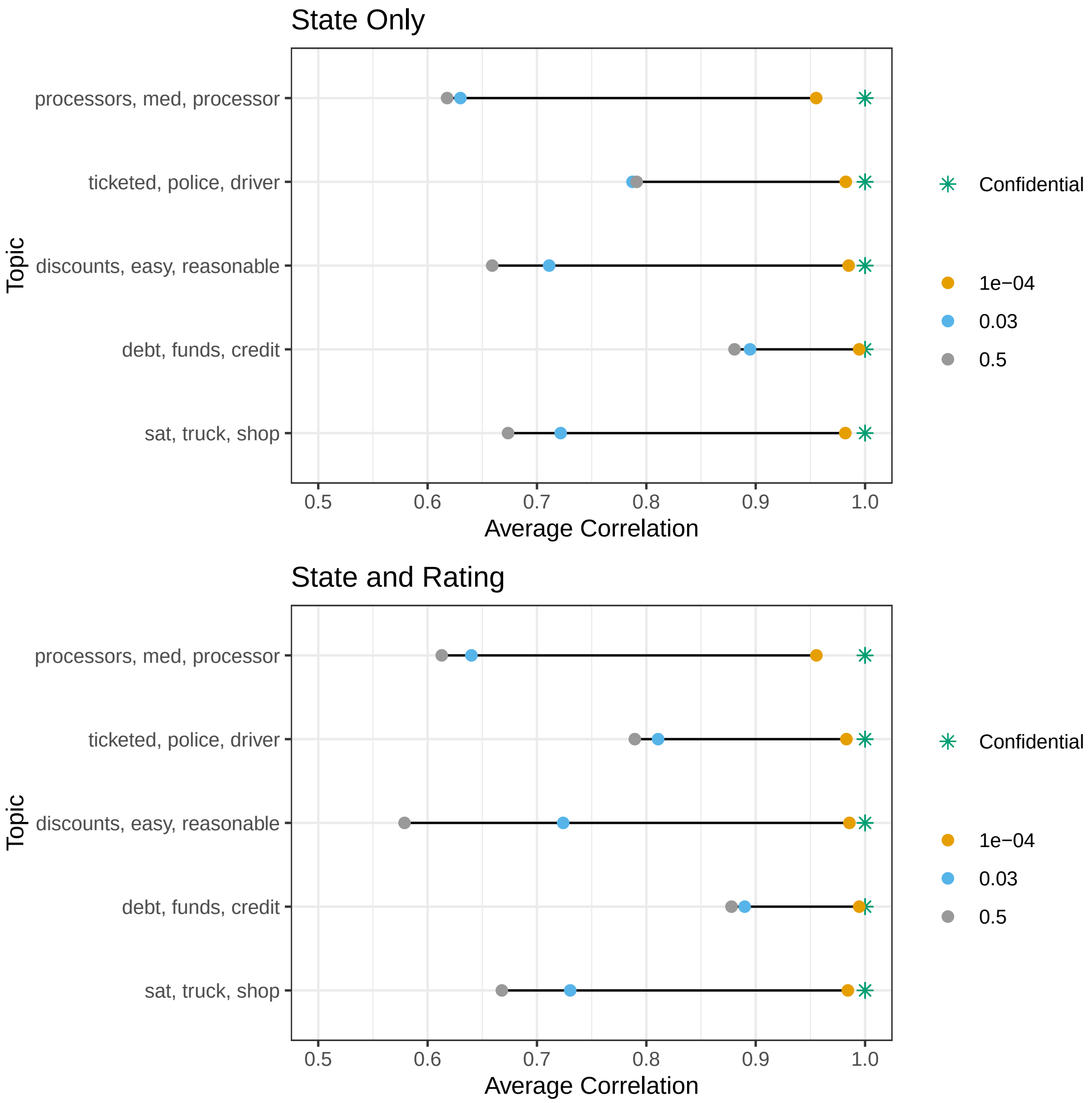


Figure 5: Average correlations of content coefficients, between confidential dataset and $M = 50$ synthetic datasets, of "State Only" synthesis and "State and Rating" synthesis.

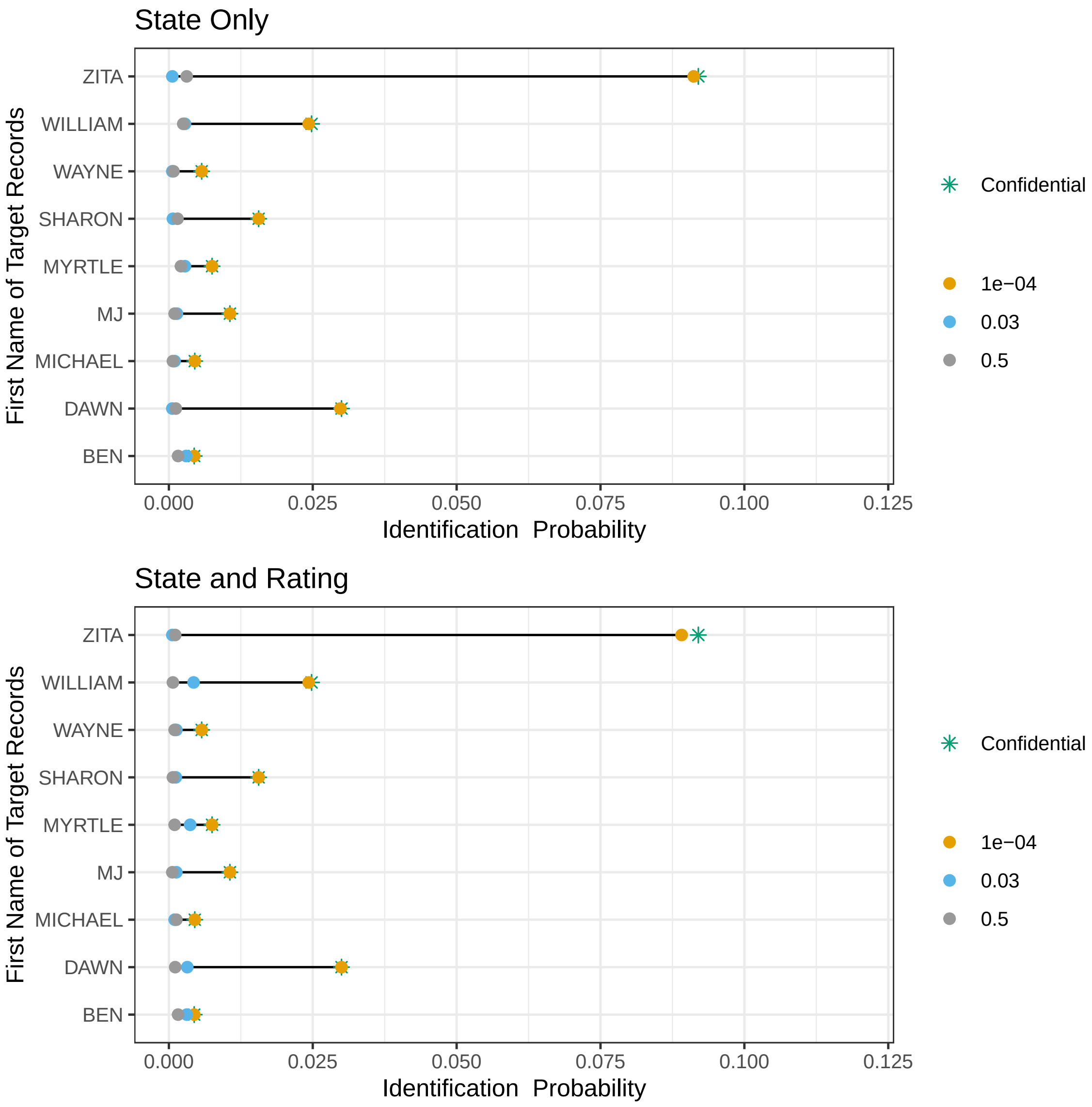


Figure 6: Identification probability for each of the 9 target records in the $M = 50$ synthetic datasets, of “State Only” synthesis and “State and Rating” synthesis.