ORIGINAL ARTICLE

Protecting the Anonymity of Online Users through Bayesian Data Synthesis

Matthew J. Schneider[1] | **Jingchen Hu**[2] | **Shawn Mankad**[3] | **Cameron D. Bale**[1]

[1]Decision Sciences, Drexel University, Philadelphia, PA, 19103, USA

[2]Mathematics and Statistics, Vassar College, Poughkeepsie, NY, 12604, USA

[3]Operations, Technology, and Information Management, Cornell University, Ithaca, NY, 14853, USA

Correspondence
Matthew J. Schneider, 3220 Market Street, Office 623, Drexel University, Philadelphia, PA, 19103, USA
Email: matt.schneider@drexel.edu

Funding information
None

User-generated content (UGC) is a modern form of communication that combines multiple types of content such as text, video, 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 Bayesian sequential synthesis strategy to generate protected structured data based on a user's adjoining textual content in their UGC. Our proposed approach enables platforms to use a single shrinkage parameter to control the privacy level of their released UGC data. We illustrate our approach on a sample of reviews of a financial services vendor from ConsumerAffairs.com. Our results show that our synthesis strategy 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.

KEYWORDS
Bayesian data synthesis, disclosure risk, 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 (Clemons, 2020). 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 protecting 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 been used and investigated, such as classification and regression trees (CART) (Wang and Reiter, 2012; Hu and Savitsky, 2018; Drechsler and Hu, 2021) and Dirichlet Process mixtures of products of multinomials (DPMPM) (Hu and Savitsky, 2018; Drechsler and Hu, 2021), 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 and evaluated for UGC data, where both structured data covariates, such as location, and textual data are present. Many of 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 DPMPM synthesizer in Hu and Savitsky (2018); Drechsler and Hu (2021)).

To tackle the challenge of protecting UGC containing both structured data covariates 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 text and 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 covariate(s) in place of the confidential one(s), 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 on the textual coefficients in the proposed synthesis model (Tibshirani, 1996). We incorporate a privacy-preserving hyperparameter, $\lambda > 0$ as the inverse scale parameter of the Laplace prior. The textual coefficients are smoothed 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 compromise user privacy. In other words, our data protection model scales the textual 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 and Feyisetan et al. (2019) used differential privacy methods to replace words with semantically similar words, neither of which is suitable for our task of protecting structured data adjoining textual data. Releasing differentially private structured data is also possible (for example, PrivBayes by Zhang et al. (2017)), but the attachment of unaltered text to it could include any information in the free text including the confidential structured data itself, breaking the differential privacy guarantee. While our proposed method faces this same issue, it still produces a highly desirable utility-risk trade-off because it is conditional on the text.

We evaluate our data protection model by analyzing the trade-off between data utility and disclosure risk (Duncan and Lambert, 1986). To evaluate global utility, we measure both the propensity score mean-squared error ratio (Snoke et al., 2018) and the relative differences in cross tabulations of categorical covariates between the confidential and the synthetic data (Drechsler and Hu, 2021). 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 offered to users by the synthetic UGC, we focus on evaluating its disclosure risk. Specifically, we calculate the probability of identification (Drechsler and Reiter, 2008; Reiter and Mitra, 2009; Hu, 2019) based on a two-stage identification strategy. 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. We compare data utility and disclosure risk results from synthetic data produced by our proposed model to results from synthetic data produced from the commonly used CART synthetic data model (Reiter, 2005).

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 using either the proposed Bayesian synthesizer or CART synthesizer to perform the sequential synthesis process decreases the probabilities of identification for users to an acceptable threshold while maintaining much of the textual content present in the structured covariates. However, important utility measures are measured at significantly better levels in synthetic data from the proposed Bayesian method, which produces a more desirable utility-risk trade-off compared to the CART.

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 risk. Section 4 is devoted to our real data application of synthesizing a sample of reviews from Consumer Affairs to demonstrate the effectiveness of our proposed synthesis model compared to the CART. We 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 consists of generating 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 textual content remains un-synthesized. Our proposed data synthesizer is a Bayesian generalized linear model for each structured data covariate to be synthesized, conditional 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, 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 by 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 until 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 $\boldsymbol{x}_i$). 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 the 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 compromise user privacy. To ensure that a high value of λ 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 (intersoft consulting, 2021).

Once the series of sequential Bayesian GLM synthesizers with shrinkage priors 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 RISK MEASURES

3.1 | Utility measures

Our utility measures consist of two types: (a) global utility measures, where we focus on evaluating whether overall distributional characteristics of the confidential data are preserved, 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

Our first measure of global utility is the propensity score mean-squared error (pMSE) ratio (Snoke et al., 2018), which we denote as $\mathcal{R}$. This is the ratio of the propensity scores calculated for the partially synthetic data to the expected propensity scores when using logistic regression to differentiate confidential and partially synthetic observations drawn from the same distribution. When synthetic data is drawn from the same distribution as the confidential data, the calculated and expected propensity scores will be approximately equal, and $\mathcal{R} \approx 1$. When $\mathcal{R} < 1$, the synthesis model is over-fit and is more likely to reproduce the confidential data, which defeats the purpose of privacy-preserving data synthesis. Over-fitting is difficult to identify using the pMSE alone since this measure indicates higher utility the closer it is to zero, where the synthetic and confidential data are identical (Snoke et al., 2018). When $\mathcal{R} > 1$, the synthesis model is under-fit and produces synthetic data with a poor approximation to the confidential data distribution.

For a second global utility measure, we note that in our setup, the synthetic covariates only include structured data covariates, all of which are categorical. Therefore, we focus on comparing the cell-level relative differences of one-way and two-way tables, constructed based on the categorical structured data covariates in the confidential dataset and in *each* synthetic dataset. Similar measures are used in Drechsler and Hu (2021) and Hu and Savitsky (2018), which have been shown as good global utility measures for synthetic data containing categorical covariates. In general, the closer the relative differences are to 0, indicating smaller deviations between the cell counts in the synthetic and confidential data, the higher the global utility. Details of both global utility measures are available in the Supplementary Materials.

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).

Prevalence 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 mth synthetic data with protection level λ. Once $M > 1$ partially synthetic datasets are generated, confidence intervals of coefficients can be calculated using widely-used combining rules (Reiter and Raghunathan, 2007; Drechsler, 2011).

Content coefficients are from a multinomial logit on words and measure how often words are used within topic t for level k_j of covariate j. We define $\boldsymbol{\beta}_{k_j} = [\boldsymbol{\beta}_{k_j}^{(1)}, \ldots, \boldsymbol{\beta}_{k_j}^{(T)}]$ 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 $\boldsymbol{\beta}_{k_j}^{\lambda,m} = [\boldsymbol{\beta}_{k_j}^{\lambda,(1),m}, \ldots, \boldsymbol{\beta}_{k_j}^{\lambda,(T),m}]$ as the $V \times T$ matrix of content coefficients computed from the mth 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 $\beta_{k_j}^{(t)}$ in the confidential data and $\beta_{k_j}^{\lambda,(t),m}$ in the synthetic data,

$$Cor(\beta_{k_j}^{(t)}, \beta_{k_j}^{\lambda,(t),m}). \tag{4}$$

This results in $K_j \times T$ 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

$$\frac{1}{M \times K_j} \sum_{k_j=1}^{K_j} \sum_{m=1}^{M} Cor(\beta_{k_j}^{(t)}, \beta_{k_j}^{\lambda,(t),m}), \tag{5}$$

which is another analysis-specific utility measure we evaluate.

3.2 | Disclosure risk 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 document/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 omit 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 risk 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 corresponding to 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,

$$\begin{aligned} Pr(ID_i = ID_z \mid \boldsymbol{D}^{\boldsymbol{\lambda}}, \boldsymbol{d}_z) = \int_{(\boldsymbol{x}_i, \boldsymbol{y}_i^{\lambda})} & Pr \quad (ID_i = ID_z \mid \boldsymbol{x}_i, \boldsymbol{y}_i^{\lambda}, \boldsymbol{D}^{\boldsymbol{\lambda}}, \boldsymbol{d}_z) \\ & Pr \quad (\boldsymbol{x}_i, \boldsymbol{y}_i^{\lambda} \mid \boldsymbol{d}_z, \boldsymbol{D}^{\boldsymbol{\lambda}}) d(\boldsymbol{x}_i, \boldsymbol{y}_i^{\lambda}), \end{aligned} \tag{6}$$

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 (6): 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,

$$Pr(ID_i = ID_z \mid \boldsymbol{D}^{\boldsymbol{\lambda}}, \boldsymbol{y}_z) = \frac{1}{|N_z^\lambda|}, \tag{7}$$

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

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

$$d(\boldsymbol{x}_i, \boldsymbol{x}_z) = \left\| \frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i}{\sigma_{\boldsymbol{x}_i}} - \frac{\boldsymbol{x}_z - \bar{\boldsymbol{x}}_z}{\sigma_{\boldsymbol{x}_z}} \right\|_1, \tag{8}$$

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 $|N_z^\lambda|$ similarities to sum to unity. The second-stage probability of identification is then

$$Pr(ID_i = ID_z \mid \boldsymbol{D}^{\boldsymbol{\lambda}}, \boldsymbol{y}_z, \boldsymbol{x}_z) = \begin{cases} \frac{s(\boldsymbol{x}_i, \boldsymbol{x}_z)}{\sum_{i' \in N_z^\lambda} s(\boldsymbol{x}_{i'}, \boldsymbol{x}_z)} & \text{if } i \in N_z^\lambda \\ 0 & \text{if } i \notin N_z^\lambda \end{cases}, \tag{9}$$

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

Our presentation and description of the identification risks evaluation focuses on the case where $M = 1$, i.e., one synthetic dataset. This is realistic for a UGC platform who ultimately selects one partially synthetic dataset to publish. 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 y_{ij} 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. This assessment can guide the selection of λ, which can then be used to generate a partially synthetic dataset for final release.

4 | APPLICATION TO CONSUMER AFFAIRS REVIEWS

4.1 | The data

We demonstrate our methodology in the substantive area of online reviews to data from ConsumerAffairs.com, 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," (ConsumerAffairs, 2018). 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 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 ! . ? # @ $ % & , ; : ( ) (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 perform synthesis and measure data utility, but all variables are used to measure the disclosure risk of our simulated synthetic data.

For our application, we define each review i having structured data $\boldsymbol{y}_i$ consisting of two covariates ($J = 2$), 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 $\boldsymbol{D} = [\boldsymbol{X}, \boldsymbol{Y}]$. 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. 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 | Synthesis scenarios and synthesis models

We consider two synthesis scenarios for the $n = 821$ reviews in the confidential dataset $\boldsymbol{D}$. The first scenario synthesizes one categorical covariate, the state, given the textual data and rating, while leaving rating un-synthesized. This scenario is labeled "State Only". The second scenario not only synthesizes state but also rating, given the textual data.

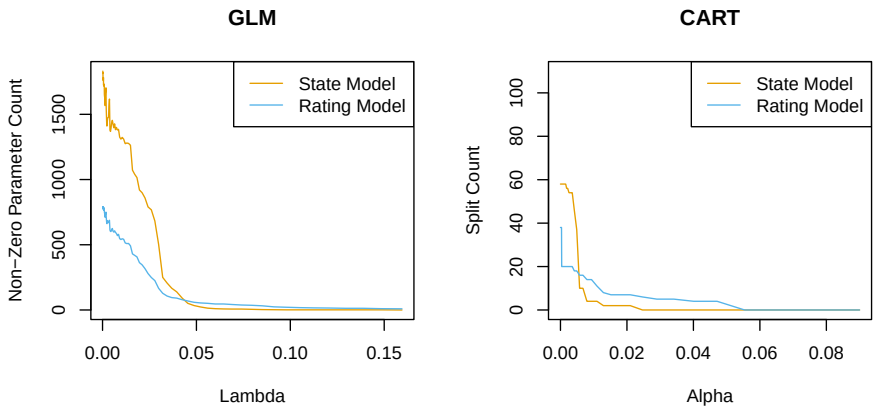

FIGURE 1 Numbers of non-zero parameters for the GLMs and splits for the CARTs used to model the State and Rating covariates.

That is, we perform sequential synthesis by synthesizing rating first and state second. This scenario is labeled "State and Rating." A detailed description of these synthesis processes is available in the Supplementary Materials.

For both synthesis scenarios, we consider two synthesis models: our GLM synthesizer proposed in Section 2 and a CART synthesizer (Reiter, 2005). Both synthesizers have a single parameter with the potential to control the utility-risk trade-off in the synthetic data. For the GLM, we use the privacy-preserving hyperparameter λ. For the CART, we use a complexity parameter $\alpha \in [0, 1]$ which requires that any split must reduce the percentage of miss-classified observations by at least α, relative to the number of miss-classified observations in the root node. Overall, smaller λ and α values should result in synthetic data with higher utility and higher disclosure risk.

A UGC platform must ultimately select a single value for the parameter λ (or α) that strikes a desirable utility-risk balance. For both synthesizers, there exists a wide range of parameter values to choose from. In Section 4.3, we propose and illustrate an approach to narrow the selection of λ to $\boldsymbol{\lambda} = \{\lambda^h, \lambda^*, \lambda^l\}$, where λ^* optimally balances utility and disclosure risk, and λ^h and λ^l correspond to *high* and *low* values of λ that should produce data with *higher* and *lower* privacy levels, all relative to λ^*. The same approach is applied to α and $\boldsymbol{\alpha} = \{\alpha^h, \alpha^*, \alpha^l\}$.

4.3 | Choosing the tuning parameters λ and α

We first estimate GLMs and CARTs for each structured covariate for a large number of potential λ and α values. We examine the number of non-zero parameters in each GLM as a function of λ and the number of splits in each CART as a function of α. This allows us to refine our choice of λ and α prior to generating synthetic data. A higher number of non-zero parameters (splits) should produce data that has high utility and also high disclosure risk. For the GLMs for both covariates, we consider a decreasing log sequence of λ values, beginning with $\lambda^{\max}$, the smallest λ for which all coefficients are zero, and ending with $\lambda^{\min}$, which is defined as a percentage of $\lambda^{\max}$. We choose $\lambda^{\min} = 0.001 \times \lambda^{\max}$. Choosing a percentage of $\lambda^{\max}$ that is too small can produce a saturated model since we have many more textual covariates than reviews. For the CARTs, we consider a log-sequence of α values from 1e−08 to 0.09. Each parameter sequence is composed of 100 values.

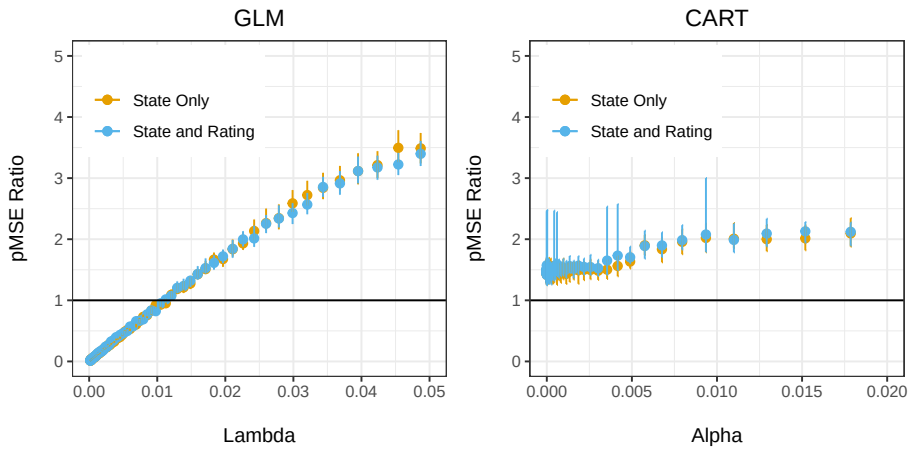

FIGURE 2 Average pMSE ratios $\mathcal{R}$ for the GLM and CART synthesized variables combined with the ten textual covariates from the random projections approach.

Figure 1 shows the numbers of non-zero parameters and splits for the synthesis models for each structured covariate and tuning parameter. Note that "State Only" synthesis will sample from a model for state (labeled as "State Model" in Figure 1), be it a GLM or a CART. Sequential synthesis, or "State and Rating", will first sample from a model for rating (labeled as "Rating Model" in Figure 1) and then sample from a model for state. The largest changes in the numbers of non-zero GLM parameters occur for $\lambda < 0.05$ (83 values out of the total 100) and, and those of CART splits occur for $\alpha < 0.02$ (90 values out of the total 100), respectively. Varying the tuning parameters within these ranges will produce the largest utility-risk trade-offs. Therefore, we restrict our last step of choosing λ and α to these ranges.

Our final step generates $M = 20$ synthetic datasets from each synthesis model with a given parameter value, λ for the GLMs and α for the CARTs. Moreover, we evaluate the pMSE ratio $\mathcal{R}$ for each synthetic dataset, introduced in Section 3.1.1 (Snoke et al., 2018). In order to calculate the pMSE of each synthetic dataset using a logistic regression, for computational feasibility, we add a dimension reduction step on the textual variables with a random projections approach that scales well for UGC platforms and has been used for textual data in other applications (Schneider and Gupta, 2016). Finally, the propensity scores are estimated with a logistic regression with main effects and interactions between all synthetic and textual covariates after dimension reduction (Snoke et al., 2018).

Figure 2 shows the means and 95% empirical confidence intervals of the average pMSE ratio $\mathcal{R}$ across 20 different random projections. Each average ratio is calculated from $M = 20$ synthetic datasets for each parameter value, i.e., for $\lambda < 0.05$ and $\alpha < 0.02$. Varying λ in the GLMs produces the expected result: Models corresponding to $\lambda < 0.01$ have a large number of non-zero parameters and are over-fit to the confidential data, whereas models corresponding to $\lambda > 0.015$ have too few non-zero parameters to accurately approximate the confidential data distribution. Varying α produces a similar effect on the fit of the CARTs, where increasing α increases $\mathcal{R}$. However, even the smallest values for α correspond to under-fit models. Moreover, there is significantly more variability in the pMSE ratios $\mathcal{R}$ across the different projections for the data synthesized from the CARTs. Depending on the randomly projected covariates used, the propensity score model is able to distinguish many of the synthetic observations within the CART synthesized data even for very small values of α.

TABLE 1 Means and 95% empirical confidence intervals for the average pMSE ratio $\mathcal{R}$ for $\boldsymbol{\lambda}$ with the GLM (left) and $\boldsymbol{\alpha}$ with the CART (right) synthesizers under "State and Rating" synthesis.

	Mean Ratio	95% C.I.		Mean Ratio	95% C.I.
$\lambda^h = 0.0487$	3.40	[3.20, 3.61]	$\alpha^h = 0.0179$	2.12	[1.89, 2.27]
$\lambda^* = 0.0112$	1.01	[0.96, 1.08]	$\alpha^* = 3.38\text{e}{-06}$	1.41	[1.26, 1.55]
$\lambda^l = 2\text{e}{-04}$	0.014	[0.013, 0.016]	$\alpha^l = 1\text{e}{-08}$	1.45	[1.28, 1.59]

TABLE 2 Average sum of absolute differences as global utility for $M = 20$ synthetic datasets for $\boldsymbol{\lambda}$ with the GLM (left) and $\boldsymbol{\alpha}$ with the CART (right) synthesizers under "State and Rating" synthesis.

	One-Way	Two-Way		One-Way	Two-Way
$\lambda^h = 0.0487$	3.00	4.69	$\alpha^h = 0.0179$	1.42	1.44
$\lambda^* = 0.0112$	1.23	1.81	$\alpha^* = 3.38\text{e}{-06}$	1.18	1.26
$\lambda^l = 2\text{e}{-04}$	0.09	0.11	$\alpha^l = 1\text{e}{-08}$	1.06	1.28

Ultimately, we use the pMSE ratio as a principled method of choosing λ^* and α^* values which correspond to models that give the best synthesis, i.e., which average $\mathcal{R}$ is closest to one, and should produce optimal utility performances. For comparison, we consider the highest (λ^h and α^h) and lowest (λ^l and α^l) values for which we estimated the average $\mathcal{R}$, giving $\boldsymbol{\lambda} = \{0.0487,\ 0.0112,\ 2\text{e}{-04}\}$ and $\boldsymbol{\alpha} = \{0.0179,\ 3.38\text{e}{-06},\ 1\text{e}{-08}\}$.

Next, we evaluate and compare the utility and disclosure risk of $M = 20$ synthetic datasets with these parameter choices. We discuss results for "State and Rating" synthesis in detail. Results for "State Only" synthesis are available in the Supplementary Materials for further reading.

4.4 | Utility results

4.4.1 | Global utility results

Measure of pMSE ratio

We report the global utility measure based on the pMSE ratio $\mathcal{R}$ for the tuning parameter choices in $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ in Table 1. The GLM with λ^* produces the highest global utility of pMSE ratio as the mean ratios are closest to one. Overall, the GLMs are more responsive to the parameter choice compared to the CARTs.

Measure of relative cell differences

Next, we present our second global utility metric 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 synthetic datasets. Figure 3 presents density plots of the relative differences of the synthetic datasets from the GLMs and CARTs. In addition, Table 2 reports the average sum of the absolute differences over $M = 20$ datasets. The relative differences for our optimal tuning parameters are similar, with the one-way average sum for λ^* and α^* only differing by about 0.05. The difference between these parameters is more pronounced for two-way tables, where α^* produces average sum of differences almost 0.60 lower than λ^*. Overall, the differences between the λ^* and α^* results are not substantial, given there are 40 cells in the two-way tables.

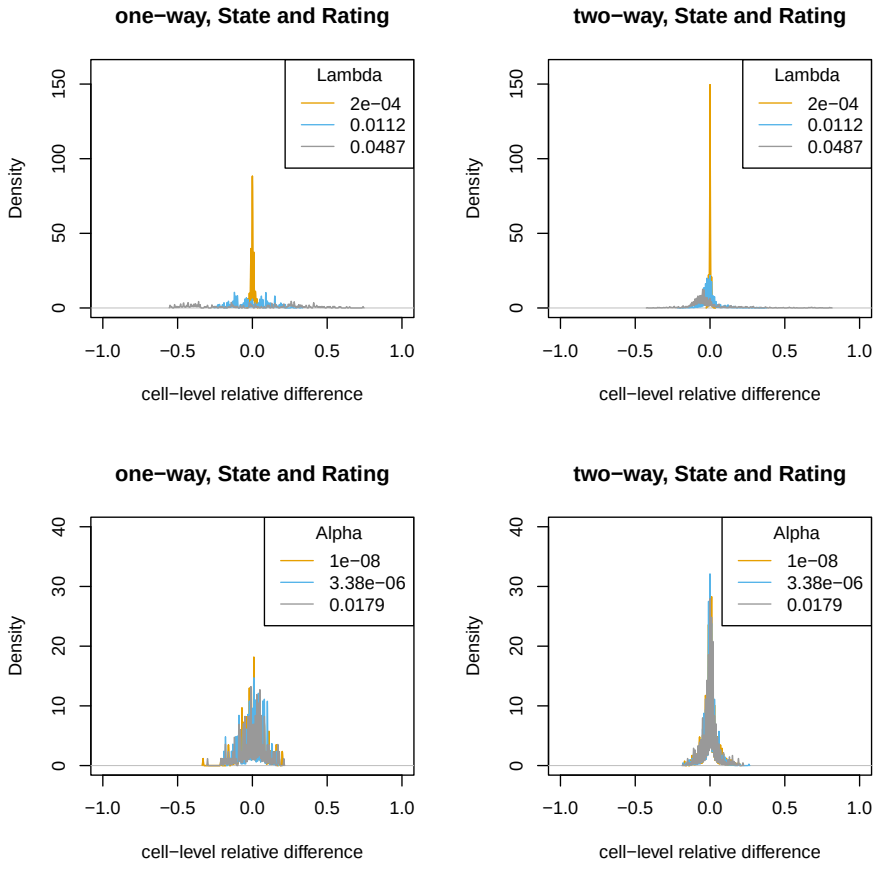

FIGURE 3 Density plots of relative differences of one-way and two-way frequencies as global utility for $M = 20$ synthetic datasets for $\boldsymbol{\lambda}$ with the GLM (Row 1) and α with the CART (Row 2) synthesizers under "State and Rating" synthesis.

4.4.2 | Analysis-specific utility results

High and low rating coefficients in the confidential data

Before presenting analysis-specific utility based on the STM introduced in Section 3.1.2, we select $T = 5$ topics and discretize rating into low rating (1 to 3 stars) and high rating (4 or 5 stars). Table 3 shows the top 3 topic words across all states, the high and low rating prevalence in California, and the low rating coefficient for all states for the *confidential* data. Topic 2 is predominantly a high rating topic for all states (e.g., the low rating coefficient is -0.582 and represents a deviation in the prevalence between high rating and low rating) and is represented by the collection of the words "rates, easy, discounts." The other four topics have positive low rating coefficients across all states and consist of mostly criticisms of the insurance product. For example, Topic 1 is composed of the words "driver, rearended, doctor" which likely describes insurance problems resulting from a traffic accident.

Measure of change in low rating coefficient

As one analysis-specific utility measure, we present the change in the low rating coefficient. Table 4 shows the low rating coefficients for Topics 1 and 2 for the GLM and CART synthesized datasets. The low rating coefficient shows moderate shrinkage for the optimal λ^*. Interestingly, we do not observe consistent shrinkage across topics for α^*. The point estimate for Topic 1 is similar to that in the confidential data, but shows shrinkage for the Topic 2 estimate similar to that observed for λ^*. For both optimal tuning parameters, the confidence intervals for Topic 1 contain the confidential estimate. The only parameter with a confidence intervals that contains the confidential Topic 2 estimate is λ^I. This shows that it is hardest to maintain the magnitudes of large coefficients, unless a UGC platform opts for low levels of privacy.

TABLE 3 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.

Topic	Topic Words	High Rating Prevalence CA	Low Rating Prevalence CA	Low Rating Coeff. (All States)
1	driver, rearended, doctor	0.125	0.279	0.154
2	rates, easy, discounts	0.685	0.103	-0.582
3	ribs, wiped, punched	-0.008	0.088	0.096
4	mileage, commissioner, storm	0.112	0.212	0.100
5	funds, deposits, deposited	0.087	0.319	0.232
Total		1.000	1.000	0.000

Measure of ratio of average low rating coefficients

We also calculate the ratio of average low rating coefficients from the synthetic data relative to the low rating coefficients estimated from the confidential data $\bar{\mu}_{k_j}^{\lambda,(t)}/\mu_{k_j}^{(t)}$, where $\bar{\mu}_{k_j}^{\lambda,(t)}$ is the average point estimate across the $M = 20$ synthetic datasets. We average these ratios across all five topics for each protection level for each synthesis model, and present these results in Table 5. On average, λ^* shrinks coefficient magnitudes by about 8% more than α^*, meaning that data users may be more likely to underestimate the prevalence of product issues related to low rating topics.

Measure of content coefficients

Our last analysis-specific utility measure utilizes the content coefficients introduced in Section 3.1.2. These measure how often all words (not only the top 3) are used in topic t for specific states. Figure 4 illustrates the average correlation of the content coefficients within states between the confidential dataset and the synthesized datasets for the GLMs and Figure 5 is for the CARTs. The average is calculated across all $K_2 = 8$ states and $M = 20$ synthetic datasets.

The GLM results show that λ^* preserves the correlations extremely well (over 0.90) for Topics 1, 2, and 5. The correlations for topics 3 and 4 are somewhat weaker, about 0.74 and 0.85 respectively. The results for the CARTs show that α^* preserves the correlations relatively well. The correlations for Topics 1, 2, and 5 are between 0.85 and 0.90, while the correlations for Topics 3 and 4 are significantly worse, about 0.60 and 0.70 respectively.

To compare the two synthesis models, we note that the correlation values for the GLM synthesized data with λ^* are significantly larger than those produced by *any* α, especially for Topics 3 and 4. These results argue strongly in favor of the proposed GLM synthesis method: Even though the CART synthesized data for α^* maintain coefficient magnitudes relatively well, the content of those topics is much less well preserved and will be more challenging for UGC platforms to draw meaningful conclusions from. For the GLM synthesized data, while the coefficient magnitudes

TABLE 4 Table of C.I.s of Topic 1 and 2 estimates of the low rating coefficient with $M = 20$ synthetic datasets for $\boldsymbol{\lambda}$ with the GLM (top) and $\boldsymbol{\alpha}$ with the CART (bottom) synthesizers under "State and Rating" synthesis.

	Topic 1		Topic 2	
	Point Estimate	**95% C.I.**	**Point Estimate**	**95% C.I.**
Conf. Data	0.154	[0.099, 0.210]	-0.582	[-0.634, -0.524]
$\lambda^h = 0.0487$	0.070	[0.014, 0.126]	-0.225	[-0.287, -0.163]
$\lambda^* = 0.0112$	0.125	[0.069, 0.182]	-0.446	[-0.504, -0.389]
$\lambda^l = 2e{-}04$	0.155	[0.101, 0.210]	-0.572	[-0.624, -0.520]
$\alpha^h = 0.0179$	0.088	[0.031, 0.145]	-0.324	[-0.384, -0.263]
$\alpha^* = 3.38e{-}06$	0.151	[0.095, 0.207]	-0.489	[-0.545, -0.434]
$\alpha^l = 1e{-}08$	0.152	[0.096, 0.208]	-0.494	[-0.549, -0.439]

TABLE 5 Table of average ratios of the low rating coefficient across $M = 20$ synthetic datasets and $T = 5$ topics, for λ with the GLM (left) and for α with the CART (right) synthesizers under "State and Rating" synthesis.

	Average Ratio		**Average Ratio**
$\lambda^h = 0.0487$	0.39	$\alpha^h = 0.0179$	0.54
$\lambda^* = 0.0112$	0.76	$\alpha^* = 3.38e{-}06$	0.84
$\lambda^l = 2e{-}04$	0.98	$\alpha^l = 1e{-}08$	0.85

under λ^* are shrunk by 24% on average, the actual topical content is well preserved, especially for Topics 1, 2, and 5.

In summary, our global utility and analysis-specific utility evaluations of "State and Rating" synthesis demonstrate that the privacy protection parameter λ in our proposed Bayesian GLM synthesizer can fine tune the utility preservation level of the resulting synthetic data. The α parameter for the CART synthesizer is unable to perform the same type of tuning, according to the analysis-specific utility measures we consider. Evaluations of "State Only" synthesis are largely the same, although we observe that utility results are often slightly worse for "State and Rating" given the extra synthesis step. We also observed that confidential point estimates of the low rating coefficients for all topics are within all the 95% confidence intervals, which is expected since "State Only" synthesis leaves Rating unchanged. We now turn to our disclosure risk evaluation results.

4.5 | Disclosure risk results

As presented in Section 3.2, for disclosure risk 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. Recall from Section 4.1 that there are 9 target reviews for our probability of identification risk evaluation. The textual content is different between the target review and the matched review, but they are written by the same author out of the total of 821 users.

Results of the second-stage identification disclosure probabilities of the 9 target records in synthetic datasets generated using the GLMs for $\boldsymbol{\lambda}$ are shown in Figure 6. Similarly, results for the CARTs for $\boldsymbol{\alpha}$ are presented in Figure 7. In addition to risk evaluation of the simulated synthetic data, we report the disclosure risk inherent in the confidential data to highlight and compare the levels of risk reduction that our synthetic data provide.

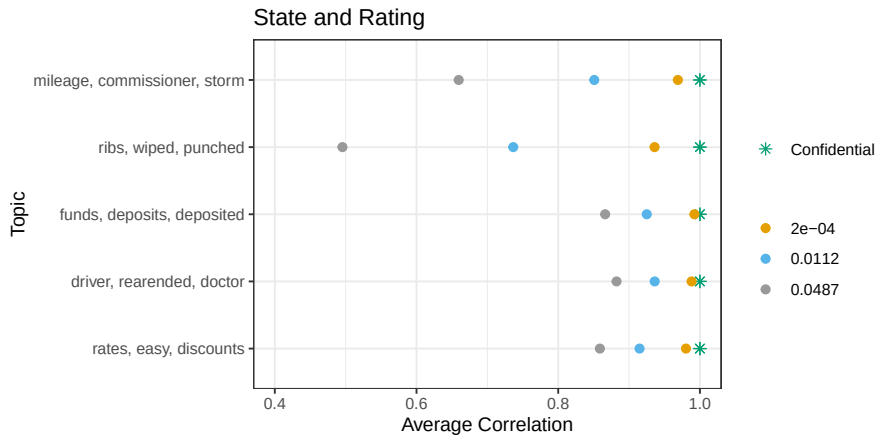

FIGURE 4 Average correlations for content coefficients between confidential dataset and $M = 20$ synthetic datasets for $\boldsymbol{\lambda}$ with the GLM synthesizer under "State and Rating" synthesis.

The green stars in Figures 6 and 7 show the disclosure risk 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 and textual content. While we illustrate the ability of our synthesizer to handle multiple categorical covariates by synthesizing State and Rating, we assume an adversary would match on State and not Rating in the first stage since Rating is review-specific. The combination of matching with the structured data and textual content results in sizeable probabilities of identification from a population of 821 users; e.g., the identification probability of "ZITA" in the confidential data $\boldsymbol{D}$ is close to 0.1.

Examining the GLM results in Figure 6, we see that synthetic data simulated with λ^* has significantly reduced identification probabilities relative to the confidential data, and the size of the decrease is related to the probability of identification in the confidential data: Individuals with higher probabilities of identification in the confidential data see a larger absolute reduction in identification probability in the synthetic data, which suggests a higher level of privacy protection for these high-risk records. For example, the identification probabilities for "ZITA" and "WILLIAM" decrease significantly, whereas the probability for "WAYNE" decreases only slightly.

These results demonstrate that our proposed Bayesian GLM synthesizer with a single privacy parameter λ fine tunes the disclosure risk of the resulting synthetic data. As λ increases with more smoothing induced in the synthetic data, the number of potential matches $|N_z^\lambda|$ increases, which leads to a lower first-stage probability of identification, $1/|N_z^\lambda|$ in Equation (7) 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.

For the CART synthesized datasets for α^*, the identification probability of each target decreases significantly relative to the probabilities in the confidential data, as shown in Figure 7. Similar to the GLM synthesized datasets, the individuals with the highest probabilities of identification in the confidential data experience the largest absolute reductions in these probabilities in the synthetic data.

To further compare the two synthesis models, we highlight that for "ZITA", the identification probability for α^* is higher than for α^I, which is the reverse of what we would hope to achieve by increasing α. By contrast, in the worst case for the GLM synthesized data, the disclosure risk for a given individual remains the same after λ is increased,

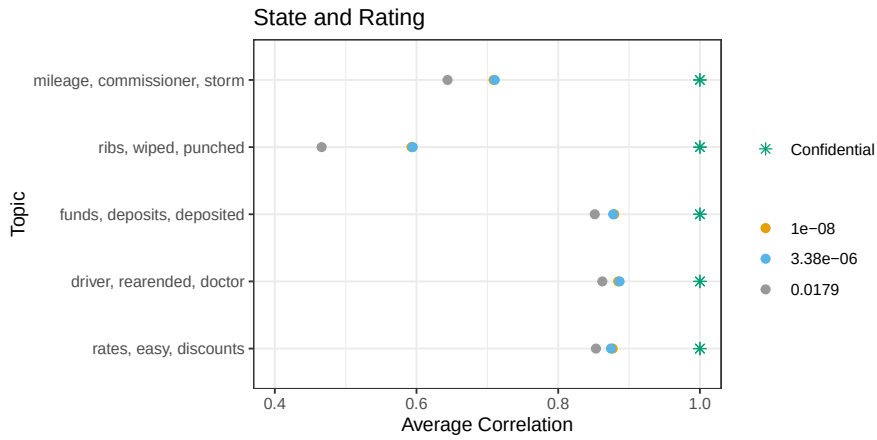

FIGURE 5 Average correlations for content coefficients between confidential dataset and $M = 20$ synthetic datasets for α with the CART synthesizer under "State and Rating" synthesis.

such as for "BEN". For some individuals, such as "WILLIAM" and "SHARON", λ^* produces reductions in disclosure risk that are comparable to those under α^*. For others, such as "DAWN" and "MJ", the disclosure risk under λ^* are slightly higher. Again, conclusions from "State Only" synthesis are largely the same. For some individuals, "State and Rating" synthesis results in slightly higher identification probabilities. While the number of potential matches $|N_z^{\lambda}|$ is similar under each synthesis type, synthesizing State conditional on synthetic Rating values appears to result in less textual similarity between reviews with the same value for State, causing the target author's writing style be more easily distinguished. This illustrates the need for protecting textual content in conjunction with structured data. Overall, the consistent results from increasing λ and comparable disclosure risk between λ^* and α^* make the GLM the preferred synthesizer.

4.6 | Concluding the synthesizer comparison

Overall, both the GLM with optimal tuning parameter λ^* and the CART with optimal tuning parameter α^* are able to produce synthetic data with relatively high utility and significantly reduced disclosure risk. However, our evaluations suggest that our proposed Bayesian GLM synthesizer has several important advantages.

First, the λ tuning parameter provides an intuitive method for directly controlling the utility-risk trade-off that is found to perform as expected across various utility and disclosure risk metrics for the GLM synthesizer. Smaller values of λ induce less smoothing, resulting in synthetic data with higher utility and higher disclosure risk. Smaller values of α also result in higher utility and higher disclosure risk, but the CARTs are much less responsive to changes in α.

Second, we find that varying λ produces a more favorable utility-risk trade-off for the GLM synthesizer. Our utility analyses show that the GLMs can generate synthetic data with much higher utility than CARTs, depending on the chosen λ. While the disclosure risk for some individuals in the CART synthesized data is lower than the GLM synthesized data, these differences are extremely small and do not justify the large reductions in analysis-specific utility in the content coefficients. As previously discussed, the optimal λ^* for the GLM provides an excellent balance of the utility-risk trade-off, whereas the optimal α^* for the CART sacrifices a large amount of analysis specific utility

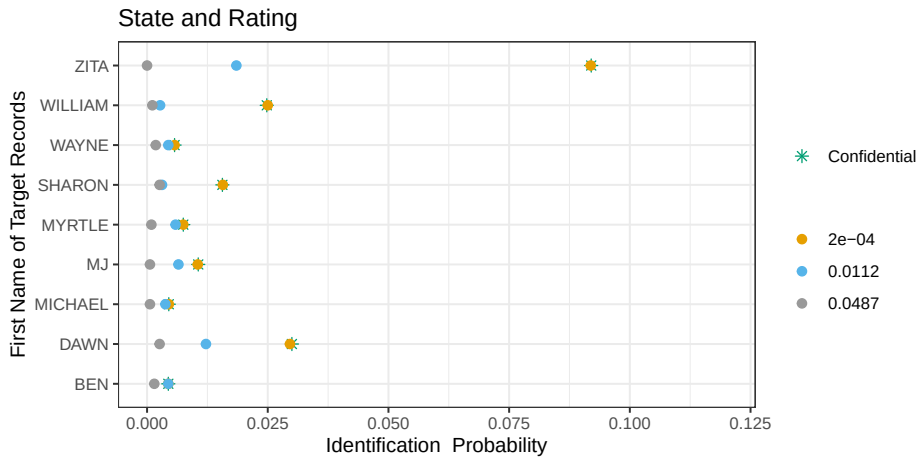

FIGURE 6 Identification probability for each of the 9 target records in the $M = 20$ synthetic datasets for $\boldsymbol{\lambda}$ with the GLM synthesizer under "State and Rating" synthesis.

in the content coefficients in favor of relatively small reductions in disclosure risk.

Furthermore, varying α produces inconsistent, and unexpected, changes in disclosure risk for the CART synthesizer. We also found that the disclosure risk for "ZITA" is in fact worse under "State Only" synthesis for α^* than under λ^* (see Figures 4 and 5 in the Supplementary Materials).

Finally, the proposed Bayesian GLM synthesizer incorporates privacy by design, since our model converges to random sampling as λ increases. By contrast, even for large values of α, a CART model will sample synthetic values according to the marginal distribution of the structured covariates, which could reveal information about the underlying data distribution.

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 risk results stabilized. We include results of $M = 50$ in the Supplementary Materials for further reading. We advocate for sufficiently large M, but not too large to avoid unnecessary computational burden.

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 levels 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

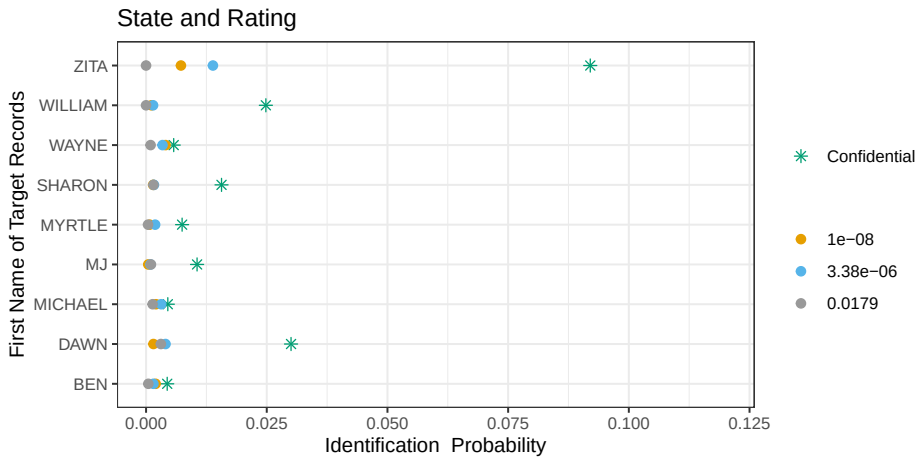

FIGURE 7 Identification probability for each of the 9 target records in the $M = 20$ synthetic datasets for α with the CART synthesizer under "State and Rating" synthesis.

amount of textual content contained in the posterior predictive density of the structured data. At moderate levels of protection, e.g., $\lambda = 0.0112$ for the Consumer Affairs data sample, the average probability of identification for an anonymous user 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 to third parties who attempt to re-identify 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 a few limitations in our paper that call for future research. 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)).

Last but not least, 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.

Clemons, C. (2020) GDPR, trust and user-generated content at work. URL: `https://www.storytagger.com/resources/gdpr-trust-and-user-generated-content-at-work/`.

ConsumerAffairs (2018) Do you know what potential buyers will find when they search for your brand online? Accessed January 7th, 2018. URL: `https://www.consumeraffairs.com/brands/#convert-detractors-2`.

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

Drechsler, J. and Hu, J. (2021) Synthesizing geocodes to facilitate access to detailed geographical information in large scale administrative data. *Journal of Survey Statistics and Methodology*, **9**, 523–548.

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-clinical-data_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.

intersoft consulting (2021) Privacy by design. URL: https://gdpr-info.eu/issues/privacy-by-design/.

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. S., 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. (2005) Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, **21**, 441.

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.

Schneider, M. J. and Gupta, S. (2016) Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, **32**, 243–256.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C. and Slavkovic, A. (2018) General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **181**, 663–688.

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*.