Data Privacy Protection and Utility Preservation through Bayesian Data Synthesis: A Case Study on Airbnb Listings

Shijie Guo and Jingchen Hu

Civil and Environmental Engineering Department, Stanford University, Stanford, CA; Mathematics and Statistics Department, Vassar College, Poughkeepsie, NY

ABSTRACT

When releasing record-level data containing sensitive information to the public, the data disseminator is responsible for protecting the privacy of every record in the dataset, simultaneously preserving important features of the data for users’ analyses. These goals can be achieved by data synthesis, where confidential data are replaced with synthetic data that are simulated based on statistical models estimated on the confidential data. In this article, we present a data synthesis case study, where synthetic values of price and the number of available days in a sample of the New York Airbnb Open Data are created for privacy protection. One sensitive variable, the number of available days of an Airbnb listing, has a large amount of zero-valued records and also truncated at the two ends. We propose a zero-inflated truncated Poisson regression model for its synthesis. We use a sequential synthesis approach to further synthesize the sensitive price variable. The resulting synthetic data are evaluated for its utility preservation and privacy protection, the latter in the form of disclosure risks. Furthermore, we propose methods to investigate how uncertainties in intruder’s knowledge would influence the identification disclosure risks of the synthetic data. In particular, we explore several realistic scenarios of uncertainties in intruder’s knowledge of available information and evaluate their impacts on the resulting identification disclosure risks.

ARTICLE HISTORY

Received September 2021
Accepted April 2022

KEYWORDS

Attributedisclosure; Data privacy; Disclosurerisk; Identificationdisclosure; Intruder’sknowledge; Syntheticdata

1. Introduction

Airbnb, one of the leading properties rental service providers, has rapidly gained popularity among people in need of lodging services, ever since its founding in 2008. An increasing number of travelers decide to choose Airbnb properties over traditional accommodations, such as hotels and hostels, when traveling (Guttentag et al. 2018). By 2020, Airbnb has over seven million listings across the globe in more than 220 countries and regions (Airbnb 2020).

However, as Airbnb customers and property hosts are enjoying the convenience it provides, concerns about the safety and security of the listed properties keep rising. Xu, Kim, and Pennington-Gray (2017) found a positive correlation between the spatial distribution of Airbnb and the number of property crimes. This result suggests that Airbnb properties may be more prone to property crimes than nonrental properties. Moreover, multiple media reports have shown that some Airbnb guests are victims of crimes including theft, sexual assault, and kidnapping, among other things (Fox-News 2015; Grind and Shifflett 2019; Bresington 2019). Thus, when releasing data about Airbnb listings to the public, the data disseminator, which could be Airbnb itself or other holders of Airbnb data, should prioritize the task of privacy protection of each released Airbnb listing. Proper protection of Airbnb listings’ sensitive information would prevent potential perpetrator from taking advantage of the released Airbnb data when committing crimes.

For the privacy protection purpose of releasing Airbnb listings data, our study considers the method of synthetic data, one of the leading approaches to providing privacy protection for the release of record-level data (Rubin 1993; Little 1993). To create synthetic data, data disseminators first build Bayesian statistical models and estimate them on the confidential record-level data. Next, synthetic values of sensitive variables are simulated from the posterior predictive distributions of the estimated Bayesian models. Data disseminators can then release the synthetic data to the public, after evaluating their utility (i.e., usefulness) and their privacy protection. We refer interested readers to Drechsler (2011) for a detailed description of the generation and evaluation of synthetic data for privacy protection. We note that government statistical agencies, such as the U.S. Census Bureau, have been creating synthetic data products for public release for more than a decade by now. For example, the OnTheMap Census product (Machanavajjhala et al. 2008) and the synthetic Longitudinal Business Database (Kinney et al. 2011), among others. There are also non-Bayesian methods for data synthesis, including classification and regression trees (CART), proposed by Reiter (2005) and can be implemented using the synthpop R package (Nowok et al. 2020).
In our case study, the New York Airbnb Open Dataset is considered as the confidential dataset to be protected (Dgomonov 2019). Among the available variables in this confidential dataset, the number of available days an Airbnb property for listing and the average listing price of an Airbnb property are deemed sensitive, and therefore to be synthesized for privacy protection. The number of available days variable, in particular, has many zero-valued records and also has a minimum of 0 days and a maximum of 365 days. To fully capture such features, we develop a Bayesian zero-inflated truncated Poisson regression model to synthesize the number of available days. As for the continuous variable of the average price of a listing, we create a Bayesian linear regression model for its synthesis. In each synthesis model, we use three predictors, the room type of an Airbnb listing, the neighborhood the listing is in, and the number of reviews the listing has on Airbnb, all of which are deemed insensitive and therefore kept unsynthesized. We build a sequential synthesis process (Kinney et al. 2011), where the number of available days variable is synthesized first, and the average price is synthesized next. In the end, we create multiple synthetic datasets with number of days and average price synthesized while the remaining variables are unsynthesized, commonly known as partially synthetic data (Little 1993).

Next, we evaluate the utility preservation performance of our simulated synthetic datasets through several data utility measures, including global utility measures such as the propensity score, the cluster analysis, and the empirical CDF; and analysis-specific utility measures such as the confidence interval overlap of quantities of interest (Woo et al. 2009; Snoke et al. 2018). We also evaluate the privacy protection performance of the partial synthetic datasets through disclosure risk evaluations, where we consider both the identification disclosure risk and the attribute disclosure risk (Hu 2019). Attribute disclosure risk refers to the risk that a data intruder correctly infers the true values of the sensitive variables in the synthetic dataset given access to external databases (Reiter, Wang, and Zhang 2014; Hu, Reiter and Wang 2014; Hu 2019; Mitra et al. 2020). Identification disclosure risk occurs when a data intruder, with certain knowledge about records in the dataset, successfully establishes a one-to-one relationship between a confidential record with itself in the synthetic dataset (Reiter and Mitra 2009). One important aspect and also an open question in the risk evaluation of synthetic data is the assumptions made about intruder's knowledge. To explore uncertainties in these assumptions, we propose and explore innovative risk evaluation approaches that incorporate such uncertainties.

Through a series of evaluations, we have found that our partial synthesis of the New York Open Airbnb Data preserves the global utility and the analysis-specific utility of the confidential data at a high level, while the attribute and identification disclosure risks of the synthetic datasets are relatively low. When we introduce uncertainties to the intruder's knowledge, the more uncertain the intruder's knowledge is, the lower the identification disclosure risks are, an expected and important result.

We note that some all-purpose synthesis methods, such as CART (Reiter 2005), are able to handle truncation and inflation of zero values that exist in the number of available days in the Airbnb sample, because CART takes random samples from the observed values in each leaf (the default settings in the synthpop R package Nowok et al. 2020). However, such a process might result in unacceptable disclosure risks, which we investigate with comparison to our proposed Bayesian synthesis models in the supplementary materials.

We now proceed to introduce the New York Airbnb Open Dataset in our case study.

1.1. The New York Airbnb Open Dataset

Our New York Airbnb Open Dataset contains 16 different variables for a total of 48,900 Airbnb listings. We focus on five most relevant variables: neighborhood, room type, number of reviews, number of available days, and the average price. Detailed descriptions of these five variables are in Table 1. For the purpose of this study, we take a random sample of \( n = 10,000 \) records from the entire dataset, ensuring records in our selected dataset have valid entries for the five variables of interest.

Among these five variables, AvailableDays and Price are considered sensitive. Therefore, we build Bayesian models to synthesize these two variables, while using the other variables as predictors. The confidential AvailableDays variable has a large number of records with zero values (its 25% quantile value is 0). Moreover, it has a minimum of 0 days and a maximum of 365 days. For modeling purpose, we apply the log transformation on the Price variable and use the resulting LogPrice variable for our case study. A table showing key summary statistics for the confidential AvailableDays and Price values is included in the supplementary materials for further reading.

The remainder of this article is organized as follows. Section 2 presents the two Bayesian synthesis models we propose and our sequential synthesis process. Section 3 evaluates the data utility of the synthetic datasets through both global and analysis-specific utility measures. Section 4 assesses the attribute and identification disclosure risks of the synthetic datasets under certain assumptions of data intruder’s knowledge. Moreover, uncertainties in some of these assumptions are explored and investigated. We end the article with a few concluding remarks in Section 5.

### Table 1. Descriptions of the five selected variables from the New York Airbnb Open Dataset.

| Variable name      | Type        | Description                                                                 | Sensitive |
|--------------------|-------------|-----------------------------------------------------------------------------|-----------|
| Neighborhood       | Categorical | The borough a listing is located in                                          | No        |
| RoomType           | Categorical | A listing’s space type                                                       | No        |
| ReviewsCount       | Count       | The number of reviews a listing received on Airbnb                          |           |
| AvailableDays      | Count       | The number of days in a year a listing is available for renting on Airbnb, with many 0s and a maximum value of 365 | Yes       |
| Price              | Continuous  | The average price of a listing for one night                                 | Yes       |
of a zero-inflated truncated Poisson regression model for AvailableDays and a linear regression model for Price. The model for Price is sequentially dependent on that for AvailableDays.

In this section, we present the details of the synthesis models in Sections 2.1 and 2.2, and describe the sequential synthesis process in Section 2.3.

2.1. Zero-Inflated Truncated Poisson Regression Model for AvailableDays

Given its features of being a count variable and having a significant number of zeros with a minimum of 0 and a maximum of 365, we develop a zero-inflated truncated Poisson regression synthesis model to synthesize the sensitive AvailableDays. AvailableDays is likely to depend on Neighborhood, RoomType, and ReviewsCount, therefore, we use three variables as predictors in the zero-inflated truncated Poisson regression model, which can be expressed as follows:

\[ Y_{\text{days,i}} \sim q_i \cdot 0 + (1 - q_i) \cdot \text{Poisson}(\lambda_i)[0,365], \]
\[ q_i \sim \text{Bernoulli}(p_i), \]
\[ \log(\lambda_i) = \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 \cdot \log X_{3i} + \epsilon_i, \]
\[ \logit(p_i) = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \cdot \log X_{3i}, \]
\[ \epsilon_i \sim \text{Normal}(0, \tau), \]

where \( Y_{\text{days,i}} \) is AvailableDays for record \( i \), \( q_i \) is binary following a Bernoulli distribution with parameter \( p_i \), and \( \lambda_i \) is the rate for the truncated Poisson distribution if AvailableDays for record \( i \) is nonzero in Equation (1). The notation Poisson(\( \lambda_i \))[0,365] refers to its truncation below 0 and above 365. As shown in Equation (3), \( \log(\lambda_i) \) is modeled given the three predictors, \( X_{1i} \) for RoomType, \( X_{2i} \) for Neighborhood, and \( X_{3i} \) for ReviewsCount, all for record \( i \). The logit of the probability \( p_i \) for AvailableDays being 0 for record \( i \) is expressed as a linear function with the same three predictors, shown in Equation (4). We give a Gamma(0.001, 0.001) prior to the parameter \( \tau \), the standard deviation of the normal error term \( \epsilon_i \) in Equation (3). We give a Normal(0, 1) prior to the hyperparameters \( \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \) and \( \beta_3 \) in Equations (3) and (4). These hyperparameters function in this hierarchical model setup to pool information across different combinations of RoomType, Neighborhood, and ReviewsCount.

2.2. Linear Regression Model for Log Price

To synthesize the sensitive Price variable, a linear regression model dependent on RoomType, Neighborhood, ReviewsCount, and AvailableDays is used to fit the log of Price in the confidential dataset. This linear regression model can be represented as the following:

\[ \log(Y_{\text{price,i}}) \sim \text{Normal}(\mu_i, \sigma_i), \]
\[ \mu_i = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 \cdot X_{3i} + \gamma_4 \cdot Y_{\text{days,i}}, \]
\[ \sigma_i \sim t(1,1)^+, \]

where \( Y_{\text{price,i}} \) is the confidential Price for record \( i \), \( \mu_i \) is the mean for the normal distribution, \( \sigma_i \) is the standard deviation of the normal distribution in Equation (6). Equation (7) shows the expression of mean \( \mu_i \), where \( X_{1i}, X_{2i}, X_{3i} \) are RoomType, Neighborhood, and ReviewsCount, respectively, while \( Y_{\text{days,i}} \) is the confidential AvailableDays for record \( i \). The notation \( t(1,1)^+ \) refers to the \( t(1,1) \) distribution on the positive real line for the standard deviation \( \sigma_i \) in Equation (8). We give a Normal(0,2) prior to the hyperparameters \( \gamma_1, \gamma_2, \gamma_3, \) and \( \gamma_4 \) in Equation (7).

Moreover, Equation (7) indicates that we sequentially synthesize \( Y_{\text{days,i}} \) first and \( Y_{\text{price,i}} \) next. We proceed to describe the details of this sequential synthesis process.

2.3. The Sequential Synthesis Process

As mentioned previously, our sequential synthesis process synthesizes AvailableDays first and Price second. Moreover, AvailableDays is used as an additional predictor for the linear regression synthesis model for log Price (Equation (7)). Such a sequential synthesis process requires fitting the synthesis models on the confidential data. Moreover, the synthesized values of variables are used in the synthesis steps of other variables where the synthesized variables serve as predictors (e.g., Kinney et al. 2011). In this section, we describe how we create \( m \) synthetic datasets and how our sequential synthesis works.

We use Just Another Gibbs Sampler (JAGS; Plummer 2003) for Markov chain Monte Carlo (MCMC) estimation of the zero-inflated truncated Poisson regression model for AvailableDays and the linear regression model for Price. To create \( m \) synthetic datasets, we generate \( m \) sets of posterior parameter draws from \( m \) independent MCMC iterations. For each set of posterior parameter draws, \( \ell = 1, \ldots, m \), we first simulate the synthetic AvailableDays, denoted as \( \tilde{Y}_{\text{days,i}}^{\ell} \) for the \( \ell \)th record in the \( \ell \)th synthetic dataset using Equation (1). Next we use the synthesized \( \tilde{Y}_{\text{days,i}}^{\ell} \) to generate the mean \( \tilde{\mu}_i^{\ell} \) for log Price using:

\[ \tilde{\mu}_i^{\ell} = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 \cdot X_{3i} + \gamma_4 \cdot \tilde{Y}_{\text{days,i}}^{\ell}, \]

where \( \gamma_1, \ldots, \gamma_4 \) refer to \( \ell \)th set of posterior parameter draws of these parameters. The resulting mean \( \tilde{\mu}_i^{\ell} \) is then used to synthesize the synthetic Price variable \( \tilde{Y}_{\text{price,i}}^{\ell} \) for the \( \ell \)th record in the \( \ell \)th synthetic dataset using Equation (6) and exponentiation. We repeat this process for every record \( i \) in the New York Airbnb Open Dataset sample and the resulting \( \ell \)th synthetic dataset contains un-synthesized \((X_{1i}, X_{2i}, X_{3i})\) as well as synthesized AvailableDays and Price \((\tilde{Y}_{\text{days,i}}^{\ell}, \tilde{Y}_{\text{price,i}}^{\ell})\).

The above process is then repeated for \( m \) times to create \( m \) synthetic datasets, \((X_{1i}, X_{2i}, X_{3i}, \tilde{Y}_{\text{days,i}}^{1}, \tilde{Y}_{\text{price,i}}^{1}), \ldots, (X_{1i}, X_{2i}, X_{3i}, \tilde{Y}_{\text{days,i}}^{m}, \tilde{Y}_{\text{price,i}}^{m})\). In our case study, we choose \( m = 20 \) to adequately capture the uncertainty in the synthesis process as well as within reasonable computation time.

3. Utility Evaluation

Data utility of synthetic data refers to how well the released synthetic data preserve the characteristics of the confidential data for analysis purposes. There are generally two types of utility measures: (a) global utility, which focuses on measuring the distance between the confidential data and the synthetic data; (b) analysis-specific utility, which emphasizes on how close
the inferential results a user would obtain from the synthetic data are to those from the confidential data (Woo et al. 2009; Snoke et al. 2018). In our investigation, we attempt both types of utility to provide a comprehensive evaluation of the utility performance of our synthesis.

Before presenting the detailed evaluation in Sections 3.1 and 3.2, we first visually inspect the synthetic data utility by plotting the synthetic AvailableDays and Price value compared to their corresponding confidential values. Figure 1(a) shows comparisons of histograms of the AvailableDays variable distribution for the confidential dataset and three synthetic datasets. In most bins in the histograms, the confidential AvailableDays has a similar count value with the synthetic values. However, in some bins, especially those toward the right tail of the distribution, the confidential AvailableDays deviates slightly from the synthetic values. Other synthetic datasets show similar results and are omitted for readability and brevity.

As for Price, Figure 1(b) shows comparisons of density plots of the confidential Price distribution and the synthetic Price distributions from three synthetic datasets. The three synthetic Price distributions are highly similar to each other, all of which largely follow the confidential distribution of the confidential Price variable despite having lower peaks. It is also worth noting that some local features in the right tail of the confidential distribution are not adequately captured by the synthetic data distributions. Other synthetic datasets show similar results and are omitted for readability and brevity.

Overall, the synthetic AvailableDays and Price distributions largely preserve the characteristics of their corresponding confidential distributions. In general, we observe a decrease in distribution similarity for Price compared to AvailableDays. This can be partly attributed to the fact that we use the synthetic AvailableDays values to sequentially synthesize the Price variable. Therefore, we expect to observe a decrease in utility for synthetic Price values in our detailed evaluation presented in the sequel.

We now proceed to the global utility and analysis-specific utility evaluations. All results are based on evaluation of $m = 20$ generated synthetic datasets.

### 3.1. Global Utility

For global utility evaluation, we consider three different measures: (a) the propensity score measure; (b) the cluster analysis utility measure; and (c) the empirical CDF measure (Woo et al. 2009; Snoke et al. 2018). All results are based on $m = 20$ synthetic datasets.

The resulting data utility measure evaluations for both the synthetic AvailableDays and Price variables are shown in Table 2, where each column corresponds to a measure: (a) Propensity $U_p$ is the propensity score utility measure (multivariate); (b) Cluster $U_c$ is the cluster analysis utility measure (multivariate); (c) eCDF-max $U_m$ and eCDF-avg $U_a$ are the maximum absolute difference and squared average difference measures of the empirical CDF utility measure (univariate), respectively.

The propensity score measure aims at distinguishing the confidential and synthetic datasets using a classification algorithm. The reported propensity score is the average of the squared distance from the estimated propensity score for each record from a constant $c$, the percentage of synthetic values in the merged dataset of the confidential and synthetic values (Woo et al. 2009; Snoke et al. 2018). In our evaluation, we calculate the propensity score measure on each synthetic dataset using a logistic regression as the classification algorithm, where the main effects of all variables are used as predictors. Since each synthetic dataset has the same size as the confidential dataset (a feature of partial synthesis; therefore, $c = 1/2$), a propensity score measure close to 0 suggests high degree of similarity between the synthetic and confidential data, which implies high data utility, whereas a propensity score measure close to 1/4 indicates low data utility. We take the average measure across the $m = 20$ synthetic datasets and the resulting value is $U_p = 0.00014$. This result suggests that the synthetic datasets are similar to the confidential dataset and thus the synthetic dataset has high data utility.

The cluster analysis measure groups the merged dataset of the confidential dataset and a synthetic dataset into multiple clusters using a clustering algorithm (Woo et al. 2009). This measure evaluates data utility based on comparing the number

| Variable | Propensity $U_p$ | Cluster $U_c$ | eCDF-max $U_m$ | eCDF-avg $U_a$ |
|----------|-----------------|--------------|----------------|----------------|
| AvailableDays | 0.00014 | 4.63 · 10⁻⁵ | 0.03140 | 0.00028 |
| Price | 0.05090 | 0.0040 | 0.00028 | 0.00028 |
| High utility | $\approx 0$ | $\approx 0$ | $\approx 0$ | $\approx 0$ |
| Low utility | $\approx 1/4$ | large $U_c$ | large $U_m$ | large $U_a$ |

Figure 1. Histograms comparing the confidential AvailableDays distribution and the synthetic AvailableDays distributions (left) and density plots comparing the confidential Price distribution and the synthetic Price distributions (right) from three synthetic datasets. The bin-width of the histograms is 25 days.
of confidential records and synthetic records in each cluster. The reported $U_c$ is the weighted sum of differences in proportions of synthetic records in each cluster and those of synthetic records in the entire merged dataset (which is 1/2 in our case study). Therefore, a $U_c$ measure close to 0 indicates the clustering algorithm cannot distinguish the synthetic data from the confidential data in each cluster, suggesting higher data utility. In our evaluation, we use the unweighted pair group method with arithmetic mean (UPGMA, Sokal 1958) method as our clustering algorithm and report the average $U_c$ value across $m = 20$ synthetic datasets. The reported $U_c = 4.63 \cdot 10^{-3}$, as shown in Table 2, is very small and close to 0. Therefore, the cluster analysis utility measure indicates high data utility of our simulated synthetic datasets.

Last but not least, the empirical CDF measure compares the empirical cumulative distribution functions (CDF) of the marginal distribution of one variable between the confidential and the synthetic data. The utility of one synthetic dataset is evaluated by whether the empirical CDF measure can discriminate the confidential dataset from this synthetic dataset (Woo et al. 2009). The empirical CDF utility has two components: (a) $U_m$, the maximum absolute difference between the empirical CDFs of the confidential dataset and the synthetic dataset; (b) $U_a$, the average squared difference between the two empirical CDFs. Both $U_m$ and $U_a$ are nonnegative. Small $U_m$ and $U_a$ values indicate high similarity between the two empirical CDFs, suggesting high synthetic data utility. Conversely, large $U_m$ and $U_a$ values indicate low data utility. We report the average $U_m$ and $U_a$ values across $m = 20$ synthetic datasets. As shown in Table 2, the $U_m$ measure values for the AvailableDays and Price variables are $U_m = 0.03140$ and $U_m = 0.05090$, respectively. The $U_a$ measure values for these two variables are $U_a = 0.00028$ and $U_a = 0.00040$, respectively. Both the maximum absolute difference and the average difference between the empirical CDFs of the confidential and synthetic datasets are close to 0, suggesting that the confidential and the synthetic data have similar CDFs, indicating high data utility.

In summary, our global utility measures of propensity score, cluster analysis, and empirical CDFs all suggest a high level of utility preservation of our synthetic datasets for the New York Airbnb Open Data sample. We next present analysis-specific utility measures on several quantities of interest.

### 3.2. Analysis-Specific Utility

To further evaluate the data utility of the synthetic datasets, we perform analysis-specific utility measures to compare a number of quantities of interest, including the mean estimate, several extreme quantile estimates, and some regression coefficient estimates. Since our synthetic datasets are partially synthetic, we obtain the point estimate and 95% confidence interval estimates of the mean and regression coefficients following the partial synthesis combining rules (Reiter 2003; Drechsler 2011). For quantile estimates, we use bootstrapping methods (Hu, Savitsky, and Williams forthcoming).

After calculating the 95% confidence interval estimates from the confidential data and the synthetic data, we use the interval overlap measure to quantify their distance (Karr et al. 2006; Snoke et al. 2018). Specifically, for the $\ell$th synthetic dataset, we calculate the interval overlap measure, $I^\ell$, for a quantity of interest, such as the mean, as the following:

$$I^\ell = \frac{U_i^\ell - L_i^\ell}{2(U_i - L_i)} + \frac{U_s^\ell - L_s^\ell}{2(U_s - L_s)},$$

where $U_c$ and $L_c$ are the upper and lower bound of the 95% confidence interval obtained from the confidential dataset, $U_i^\ell$ and $L_i^\ell$ are the upper and lower bound of the 95% confidence interval obtained from the $\ell$th synthetic dataset. Moreover, $U_s^\ell = \min(U_i^\ell, U_s^\ell)$, and $L_s^\ell = \max(L_i^\ell, L_s^\ell)$, where subscript $i$ refers to intersection (of the two confidence intervals). Finally, we report the average $I = 1/m \sum_{\ell=1}^m I^\ell$ as the interval overlap for the quantity of interest. An $I$ value close to 1 indicates a high degree of overlap, suggesting high data utility of the synthetic datasets. In contrast, an $I$ value close to 0 or even negative indicates little or no overlap, suggesting poor data utility of the synthetic datasets.

We first consider the point estimate, the 95% confidence interval, and the interval overlap measures for AvailableDays and Price, where we report those for the mean, the 25% quantile, and the 90% quantile. Results for the 90% quantile are included in Figure 2 whereas the other ones are included in the supplementary materials for brevity. In Figure 2, panel A is for AvailableDays and panel B is for Price. In each panel, we include the point estimate (box with a cross) and the 95% confidence interval (the vertical bars) of the confidential data (black and left) and of the synthetic data (yellow and right). Moreover, we add the interval overlap measure $I$ for each quantity in the panel.

Overall, with the exception for the mean of AvailableDays, the point estimates of the three quantities of interest are close between the confidential data and the synthetic data for AvailableDays, and the 95% confidence interval overlap measures are close to 1 as well, suggesting a high level of utility preservation. For the 25% and 90% quantile estimates, the results for Price are slightly worse, which are expected since Price is synthesized after AvailableDays is synthesized according to the sequential synthesis process. These results suggest that when using the synthetic data, users would obtain point estimates and 95% confidence interval estimates for key quantities of AvailableDays and Price that are overall close to those from the confidential data.

We also consider regression analysis types of analysis-specific utility measures. Specifically, we examine the regression coefficients of the effects of Neighborhood, RoomType, ReviewsCount, and AvailableDays on the response variable, Price, using a linear regression. Recall that both AvailableDays and Price are synthesized, therefore, this regression analysis utility measure takes into account all synthesized variables. As with the mean estimates, partial synthesis combining rules are applied in obtaining the regression coefficients from the $m = 20$ synthetic datasets. For each coefficient, we have a point estimate and a 95% confidence interval estimate. Given the confidence interval estimates from the confidential data and the synthetic data, we calculate the interval overlap measure. Selected results are displayed in Figure 3.

The point estimates of the coefficients for ReviewsCount and AvailableDays given the synthetic data are close to those given the confidential data. In both cases, the interval overlap measures are almost 1, indicating an extremely high level of utility.
Results for the other regression coefficients show similarly high utility with interval overlap measures above 0.9 for most of them. These are omitted for brevity.

In summary, our analysis-specific utility measures on key quantities of interest, such as means, quantiles, and regression coefficients, all suggest a high level of utility preservation of the generated synthetic datasets for the New York Airbnb Open Data sample. Combined with the results of global utility evaluation in Section 3.1, we conclude that the synthetic datasets generated from our proposed sequential synthesis models are able to preserve important characteristics of the confidential New York Airbnb Open Data sample. We now proceed to evaluate the disclosure risks associated with these synthetic data.

4. Disclosure Risk Evaluation

The utility evaluation results presented in Section 3 demonstrate a high level of utility preservation of our synthetic datasets. However, synthetic data should only be considered for public release if their level of privacy protection is satisfactory. In this section, we assess the disclosure risks associated with our simulated synthetic datasets. Since our synthesis is partial synthesis of AvailableDays and Price, we consider attribute disclosure risk and identification disclosure risk, both of which exist in partially synthetic data (Hu 2019).

4.1. Attribute Disclosure Risk

Attribute disclosure risk refers to the risk that a data intruder correctly infers the confidential values of sensitive, synthesized variables given the released synthetic dataset and other information from external databases. In our case study, AvailableDays and Price are synthesized whereas RoomType, Neighborhood, and ReviewsCount are kept unsynthesized. If an intruder has access to information on records’ information on some or all of these three unsynthesized variables, attribute disclosure could occur.

In our evaluation, we measure the attribute disclosure risk by estimating the sum of the average probabilities that the attributes of the sensitive variables are exposed for each record in the confidential dataset. Our approach is similar to the one proposed by Mitra et al. (2020). The process is applied to each of the m synthetic datasets and we drop the synthetic dataset index ℓ for notation simplicity in this section. Once the probabilities are calculated for synthetic dataset ℓ, we calculate and report the average across m synthetic datasets as the attribute disclosure risk measure. Results are based on m = 20 synthetic datasets.

For record i, we first identify all records in the synthetic dataset that are matched with record i based on the unsynthesized variables. Denote this set of matched records as \( \mathcal{M}_i \). Next, we calculate the probability that the attributes of record i are exposed to the intruder, which is defined as the fraction of records that share similar sensitive attributes among the matched set of records in \( \mathcal{M}_i \). In our case study, we use the unsynthesized RoomType, Neighborhood, and ReviewsCount for the matching step and consider similar sensitive attributes for synthesized AvailableDays and Price. Specifically, we consider records in \( \mathcal{M}_i \) that have AvailableDays within ±5 or ±10 days and Price within ±5% or ±10% of Price of record i as similar, following the choices made in Mitra et al. (2020).

Mathematically, the attribute disclosure risk probability of record i can be expressed as the following:
where \( I() \) is the indicator function that takes value 1 if the condition is met and 0 otherwise, \( c_i \) is the number of records in \( M_i \), \( r_{\text{avail}} \) and \( r_{\text{price}} \) are the radius within which records are considered similar, and finally \( p_i \) represents the attribute disclosure risk probability for record \( i \). In our case study, we consider the absolute difference between AvailableDays with a radius of \( r_{\text{avail}} \in \{5, 10\} \) and the percentage absolute difference between Price with a radius of \( r_{\text{price}} \in \{5\%, 10\\%\} \).

We denote the final attribute disclosure risk measure of the entire synthetic dataset by \( AR = \sum_{i=1}^{n} p_i \), and we compare and report the attribute disclosure risk of the synthetic and confidential datasets under different \( r_{\text{avail}} \) and \( r_{\text{price}} \) choices, shown in Table 3.

As evident in the last column of Table 3, the attribute disclosure risk of the synthetic datasets increases as the radius choices increase. This result is intuitive since a more stringent attribute similarity definition (i.e., smaller radius values) will result in a lower probability of finding similar attributes in the synthetic data. This result is intuitive, since a more stringent attribute similarity definition will result in a lower probability of finding similar attributes in the synthetic data.

### 4.2. Identification Disclosure Risk

Identification disclosure occurs when a data intruder, with certain knowledge, correctly establishes a one-to-one relationship between a record in the confidential data and this record in the synthetic data. Given the nature of our synthesized variables, count for AvailableDays and continuous for Price, we follow the approaches in Hornby and Hu (2021) when declaring a match between a record in the confidential data with another record in the synthetic data. Specifically, among all records sharing the same unsynthesized RoomType, Neighborhood, and ReviewsCount in the confidential and synthetic data, a match is declared for confidential record \( i \) if for a synthetic record \( j \), its synthetic AvailableDays value is within 5 days from the confidential AvailableDays value of record \( i \) and if the log transformation of its synthetic Price value is within the ±5% range of the confidential log (Price) value of record \( i \). The choices of ±5 days for AvailableDays and ±5% for Price are based on previous studies (Mittra et al., 2020) and the Airbnb data sample. A thorough investigation of the choices would be an interesting future research direction. As before, evaluation is performed on each synthetic dataset separately and results are combined by taking averages. All results are based on \( m = 20 \) synthetic datasets.

The identification disclosure risks of the confidential and the synthetic data are evaluated based on three commonly-used measures, namely the expected match risk (EMR), true match rate (TMR), and false match rate (FMR) (Reiter and Mitra, 2009; Hu, 2019; Drechsler and Hu, 2021), presented in Table 4.

| r_{\text{avail}} | r_{\text{price}} | Confidential AR | Synthetic AR |
|------------------|------------------|-----------------|--------------|
| 5                | 5%               | 636.40          | 126.07       |
| 10               | 5%               | 657.53          | 146.44       |
| 10               | 10%              | 816.55          | 291.73       |

The EMR represents the expected number of records in the synthetic datasets to be correctly identified. The TMR refers to the percentage of true unique matches among all records, and the FMR measures the percentage of false unique matches, respectively, among all unique matches. Therefore, lower EMR, lower TMR, and higher FMR indicate lower identification risks, and vice versa.

Overall results in Table 4 suggest low identification risks associated with our synthetic datasets for the New York Airbnb Open Data sample. Indeed, compared to the inherent identification disclosure risk of the confidential data, the synthetic data show a striking reduction in the probability that a record will be correctly identified by data intruder by more than 50-fold (i.e., 0.7182 vs. 0.0126, as calculated by the values in the EMR column in Table 4 divided by \( n = 10,000 \) in the sample). This evaluation is based on the assumption that the intruder would have perfect knowledge of the un-synthesized variables, RoomType, Neighborhood, and ReviewsCount, which could be unrealistic in practice and potentially overstate the disclosure risk concerns. To evaluate the impact of uncertainties in intruder's knowledge, we proceed to incorporate scenarios of such uncertainties in the next section.

### 4.3. Uncertainties in Intruder's Knowledge in Identification Disclosure Risk Evaluation

To investigate how uncertainties in intruder's knowledge would influence the identification disclosure risks of the synthetic data, we consider adding normal noise to the ReviewsCount variable after applying a log transformation. This is to mimic uncertainties in intruder's knowledge about this ReviewsCount variable. Specifically, for record \( i \),

\[
\log(\text{ReviewsCount}_i^*) \sim \text{Normal}(\log(\text{ReviewsCount}_i), S \times \text{ReviewsCount}_i),
\]

where ReviewsCount\(_i^*\) represents the intruder's knowledge of the ReviewsCount value, ReviewsCount\(_i\) is the confidential ReviewsCount value, and \( S \times \text{ReviewsCount} \) represents the level of uncertainties in intruder's knowledge, which is the standard deviation.

| AR | EMR      | TMR  | FMR  | u  |
|----|----------|------|------|----|
| Confidential IR | 7182.03 | 0.62 | 0.00 | 6226 |
| Synthetic IR    | 125.59  | 0.01 | 0.91 | 1153 |
deviation of the normally added noise. We experiment with the percentage $S$ from $\{0.00, 0.01, \ldots, 0.15\}$ and evaluate the identification disclosure risks under each different $S$ value as the uncertainty level of ReviewsCount. As before, results of EMR, TMR, and FMR are reported with each $S$ value. We also report the number of unique matches in each case.

Results are presented in Figure 4 for all different experimented $S$ values. Evidently, as the level of uncertainties increases (i.e., $S$ increases), both EMR and TMR decrease, while FMR increases. Therefore, in general, as data intruder’s knowledge becomes more uncertain, the identification disclosure risks for the synthetic datasets decrease, a not-so-surprising result. These can be explained by the fact that the uncertainties in ReviewsCount make it more difficult for the intruder to establish a one-to-one relationship between two records, one in the confidential data and the other in the synthetic data. However, for the same reason, uncertainties in intruder’s knowledge would reduce the inherent identification disclosure risk of the confidential dataset. Nevertheless, for all uncertainty levels $S$ we considered, the synthetic data offer significant improvement in privacy protection by reducing the probability of identifying a record by at least 50-fold. Results are included in the supplementary materials for further reading.

It is worth noting that while panel D in Figure 4 shows that in general the number of unique matches in the synthetic datasets increases as $S$ increases, the FMR rises rapidly with as the amount of uncertainties in intruder knowledge increase. These results suggest that uncertainties in ReviewsCount increase the probability of a unique match to be a false match, and higher uncertainty level makes it harder for data intruder to correctly identify records in the confidential data with the released synthetic data.

5. Conclusion

As synthetic data become more widely used by data disseminators, our case study focuses on how to best disseminate a sample of the New York Airbnb Open Data, where the number of available days and the price of Airbnb listings are considered sensitive information to be protected. Given the features of the number of available days, we propose a zero-inflated truncated Poisson regression model. Moreover, we use a sequential synthesis approach to further synthesize the price variable after synthesizing the available days. The resulting multiple partially synthetic datasets are evaluated for their utility and privacy protection performances.

Multiple utility measures, including global utility and analysis-specific utility, are evaluated for our synthetic New York Airbnb Open Data samples. Overall, our synthetic datasets preserve a high level of utility that can be hugely beneficial for users when conducting analyses. Attribute disclosure risk and identification disclosure risk are evaluated for our synthetic datasets under certain intruder’s knowledge assumptions. Overall, there exist low disclosure risks in our synthetic datasets, making them suitable for public release.

An important methodological contribution of our work lies in our investigation of uncertainties in intruder’s knowledge when evaluating identification disclosure risks. Specifically, we consider how the uncertainty level of intruder’s knowledge of an unsynthesized variable impacts several summaries.
of identification disclosure risks. Our results show that as intruder’s knowledge becomes more inaccurate (i.e., more uncertain), it is less likely for the intruder to correctly identify records in the confidential data with access to the released synthetic data and external information.

Extending the uncertainty evaluation to attribute disclosure risks is an important and highly relevant future work direction. As with identification disclosure risk evaluation, assumptions are made when evaluating attribute disclosure risks. Investigating the impact of intruder’s knowledge uncertainty in attribute disclosure risks would shed light on how to most realistically evaluate disclosure risks in synthetic data.

**Supplementary Materials**

Readers are referred to the Supplementary Materials for a detailed comparison of data utility preservation and disclosure risk protection between the CART synthesis method and the proposed Bayesian synthesis model. Additional information and results from Section 1.1, Section 3.2, and Section 4.3 are also included in the Supplementary Materials for further reading.

**Acknowledgments**

The authors thank Joerg Drechsler for his expertise on the CART synthesis method.

**ORCID**

Shijie Guo  
https://orcid.org/0000-0002-9199-1502

**References**

Airbnb (2020), “About Us,” Available at https://news.airbnb.com/about-us/. Accessed August 03, 2020. [192]

Bressington, B. (2019), “How Airbnb Hosts are Joining Crime Networks to Steal from Guests,” available at https://medium.com/@BenBressington/how-airbnb-hosts-are-joining-crime-networks-to-steal-from-guest-edc4a035c3#.f [192]

Dgononov (2019), “New York City Airbnb Open Data,” available at https://www.kaggle.com/dagononov/new-york-city-airbnb-open-data. Accessed April 20, 2020. [193]

Drechsler, J. (2011), Synthetic Datasets for Statistical Disclosure Control, New York: Springer. [192,196]

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. [198]

Fox-News (2015), “10 Incredible Airbnb Horror Stories,” available at https://www.foxnews.com/travel/10-incredible-airbnb-horror-stories. [192]

Grind, K., and Shifflett, S. (2019), “Shooting, Sex Crime and Theft: Airbnb Takes Halting Steps to Protect its Users,” available at https://www.wsj. com/articles/shooting-sex-crime-and-theft-airbnb-takes-halting-steps-to-police-its-platform-11577374845. [192]

Guttentag, D., Smith, S., Potwarka, L. and Havitz, M. (2018), “Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study,” Journal of Travel Research, 57, 342–359. [192]

Hornby, R., and Hu, J. (2021), “Identification Risks Evaluation of Partially Synthetic Data with the IdentificationRiskCalculation R Package,” Transactions on Data Privacy, 14, 37–52. [198]

Hu, J. (2019), “Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data,” Transactions on Data Privacy, 12, 61–89. [193,197,198]

Hu, J., Reiter, J. P., and Wang, Q. (2014), “Disclosure Risk Evaluation for Fully Synthetic Categorical Data,” in International Conference on Privacy in Statistical Databases, pp. 185–199, Springer. [193]

Hu, J., Savitsky, T. D., and Williams, M. R. (forthcoming). “Risk-Efficient Bayesian Pseudo Posterior Data Synthesis for Privacy Protection,” Journal of Survey Statistics and Methodology. [196]

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006), “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality,” The American Statistician, 60, 224–232. [196]

Kinney, S. K., Reiter, J. P., Resnek, 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, 363–384. [192,193,194]

Little, R. J. A. (1993), “Statistical Analysis of Masked Data,” Journal of Official Statistics, 9, 407–426. [192,193]

Machanavajhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008) “Privacy: Theory Meets Practice on the Map,” in Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, 277–286. IEEE Computer Society. [192]

Mitra, R., Blanchard, S., Dove, I., Tudor, C., and Spicer, K. (2020), “Confidentiality Challenges in Releasing Longitudinally Linked Data,” Transatlantic Data Privacy, 13, 151–170. [193,197,198]

Nowok, B., Raab, G. M., Dibben, C., Snoke, J., and van Lissa, C. (2020), sftophop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control, R package version 1.6-0. [192,193]

Plummer, M. (2003), “JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling,” in Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vol. 124, pp. 1–10. [194]

Reiter, J. P. (2003), “Inference for Partially Synthetic, Public Use Microdata Sets,” Survey Methodology, 29, 181–188. [196]

Reiter, J. P., and Mitra, R. (2009), “Estimating Risks of Identification Disclosure in Partially Synthetic Data,” Journal of Privacy and Confidentiality, 1, 99–110. [193,198]

Rubin, D. B. (1993), “Discussion Statistical Disclosure Limitation,” Journal of Official Statistics, 9, 461–468. [192]

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, 181, 663–688. [193,195,196]

Sokal, R. R. (1958), “A Statistical Method for Evaluating Systematic Relationships,” The University of Kansas Science Bulletin, 38, 1409–1438. [196]

Woo, M.-J., Reiter, J. P., Oganian, A., and Karr, A. E. (2009), “Global Measures of Data Utility for Microdata Masked for Disclosure Limitation,” Journal of Privacy and Confidentiality, 1, 111–124. [193,195,196]

Xu, Y.-H., Kim, J.-w., and Pennington-Gray, L. (2017), “Explore the Spatial Relationship between Airbnb Rental and Crime,” in Travel and Tourism Research Association: Advancing Tourism Research Globally, Vol. 5. [192]