Synthetic Differential Privacy Data Generation for Revealing Bias Modelling Risks

Matthew Wilchek
M.S. Data Science
George Washington University
Washington, D.C., USA
mwilchek@gwu.edu

Yingjie Wang
M.S. Data Science & Analytics
Georgetown University
Washington, D.C., USA
yw592@georgetown.edu

***Abstract*— Personally identifiable information (PII) continues to be used in predictive modeling by academic researchers and industry organizations. Most notably, the healthcare industry has been a popular testbed for innovative approaches from academia and institutions to address research using PII in predictive applications and synthetic data generation. The majority of these approaches that generate synthetic PII are based on actual data or obfuscating real data parts. Privacy leakage and ethical disclosure results continue to be among the largest issues that are difficult to avoid in synthetic PII generation techniques. In this analysis, we propose a novel method to generate synthetic, differential privacy data while avoiding the common pitfalls and capable of being leveraged broadly. Evidence is also shown that proves how our novel approach can maintain inference for modeling and potential risks tied to PII features. We conclude with a summarization of our findings and results and a short discussion on how using PII data may impact organizations interested in developing predictive applications.**

Keywords*— *Machine Learning, Synthetic Data, Data Ethics, Privacy, Disclosure, Anonymity

I. Introduction

Synthetic data is no longer a new concept of generating a replica of real data to prevent privacy leakage or as an easier means to create data where the collection process is not feasible in a specific context. In academia and throughout the industry, the desire to generate replica data continues to grow. Synthetic data can now potentially be in various forms such as images, numerical or categorical. The two primary challenges around using synthetic data for predictive modeling are veracity and value.

The accuracy of replica data is critical so that it represents the actual data desired for analysis or predictive modeling. Implementing this accuracy can be particularly difficult depending on the type of synthetic data that needs to be generated. A prime example and focus of this paper is Personally Identifiable Information (PII). Researchers have suggested a couple of ways PII could be developed to be as accurate as a real PII dataset, such as a Generative Adversarial Network (GAN). GAN is a machine learning model that can generate new data instances that resemble a training dataset. However, specific industries cannot reveal their training datasets due to possible privacy leakage picked up by the GAN, such as healthcare, social media, or law enforcement.

Obtaining an authentic replica of a real dataset would also mean that that dataset's same value would be maintained. The same trends, insights, and anomalies should still be observed in the replica data. Copying those significances and small details around PII, though, may prove more challenging.

Synthetic health data is an excellent area of synthetic PII research. Realistic patient data can be challenging to access because of cost, patient privacy concerns, or other legal restrictions. It can also be a vital solution for improving the measurement and enhancement of the value of care for future patients. In the past few years, there have been several promising new research methods on generating synthetic PII for healthcare and what it could mean in bioinformatics.

Another industry not discussed as much for how PII data can be generated or used in predictive modeling applications, is the travel industry. Many travel companies, immigration institutions, advertising services, and even law enforcement agencies are interested in collecting PII. Predicting PII features about an individual for where and when they will travel can be compelling intelligence for both profitable and defensive means. PII collected for this application would certainly be significant for accurate applications and susceptible to several privacy concerns and legal restrictions.

One of the biggest benefits PII can bring for training a machine learning model is the strength and feature significance it has when it's predicting almost anything about people, such as interests, behavior, or backgrounds [10]. Netflix, for example, has been praised for its intelligent recommendation algorithms. A core data component behind their successful algorithm is the behaviors of the users on their platform and the personal interests they have in what they want to watch. As a user on Netflix, one of the first questions asked upon account creation is the permission for Netflix to recommend the user content that they might like based on their recent activity. However, some may not realize that they are also approving Netflix can or cannot collect data on them. Permission to use collected PII data for predictive applications is perhaps the biggest challenge for any institution due to privacy rights. Depending on what kind of industry is interested in using PII, some privacy laws may be more challenging to deal with. This is especially where synthetic PII data can look alluring if it can circumvent privacy concerns.

This article proposes a novel method of creating synthetic PII that could resemble a real-world travel dataset used to predict features for an individual based on where they may travel to. To add context to the problem, we take the perspective of what PII may be available for migrants traveling from Mexico to the United States. Migration movements have been a hot topic in the news for the United States. They are likely an issue targeted by migrant and policy research institutions or even law enforcement agencies. Often

many socioeconomic factors contribute to migrant travel; however, if PII is collected on these individuals, it could likely strengthen a predictive application.

We also look at how the PII features statistically compare when used to predict where a migrant might appear in the United States and how many days it may take them. We apply several statistical tests to the synthetic PII data, such as contingency tables, correlation tests using Cramer's V and Theil's U, and unsupervised modeling results using association rule mining.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work, and Section 3 details the proposed generation of synthetic differential privacy data. The data's validating statistical inference is described in Section 4, followed by experimental modeling results in Section 5. Section 6 and 7 conclude the analysis and discuss why greater coordination and partnership may need to occur between academia and privacy protection institutions to ensure enough regulation and caution is met within organizations using PII data for predictive purposes. Ethical risks for using PII features are also proposed if ever publicly disclosed. The full synthetic dataset and scripts will also be available through this article.

II. Related Work

This section provides a review of the latest developments in creating synthetic datasets related to PII, followed by a legal and ethical perspective on how PII could be used in predictive modelling from industry or academic institutions.

A. Synthetic PII Development

One of the more notable articles on creating synthetic data related to PII was electronic healthcare records (EHR) [17]. Creating synthetic EHR can be a sensitive process due to the potentially high penalties in the accidental release of actual patient data. It is also possible that health histories recovered from obfuscated data could result in discrimination in any analyses.

As a result of Yale's et al. (2019) research, they developed a highly effective Wasserstein Generative Adversarial Network (WGAN) trained from the dataset MIMIC-III (Medical Information Mart for Intensive Care), which consists of de-identified ICU (intensive care unit) data from 2001 to 2012 [1]. Most of the PII data from MIMIC-III consists of patient identification numbers, gender, date of birth, and relevant patient data such as prescriptions or medical events tied to them. Ultimately, the WGAN was evaluated and optimally measured of differential privacy (DP) metric. It was concluded that the WGAN, later renamed "HealthGAN," was the most effective method that could maintain privacy and allow model export. However, all GANs trained and evaluated from the MIMIC-III dataset still resulted in some level of measurement in the DP metric or their copy of original data (CP) metric, which could mean an actual record from the original dataset.

Yale et al. (2019) later details in the following article additional new metrics to quantify the susceptibility of GAN models that could leak information about the individual data records on which they were trained. Adversarial machine learning is becoming more of a concern for researchers to consider in training a model, especially those that include PII. For a target model to leak information, a separate machine learning inference model can be trained to recognize differences in a target model's predictions on the inputs that it trained on compared to the inputs it did not train on [15]. Therefore, while HealthGAN may be a possible option to create synthetic PII with a low chance of being susceptible to an adversarial machine learning attack, the need to still research the development of secure, synthetic PII still exists today. In addition, HealthGAN is also focused on training from synthetic PII related to healthcare data instead of other industry-specific datasets.

Another resource of data accessible to academia and industry for predictive modelling is from government agencies. Federal government agencies often most follow a strict protocol to publish sensitive or internal datasets. Often, these processes involve data wrangling and a variety of legal steps to ensure nothing classified, sensitive, or unwanted privacy disclosure is divulged. As such, government agencies are also interested in creating "fully synthetic" or "partially synthetic" data from the original datasets [13].

Fully synthetic data can protect confidentiality since the released data are not trained from original records and do not contain any collected values of the real world. Through appropriate fully synthetic data generation, analysts should still be able to make valid inferences for various measurements using standard statistical methods and software. Other valuable features of fully synthetic data are described by [4], [7], [11], [14], and [19].

Partially synthetic data includes records initially collected from the source, but any original PII values are replaced with multiple imputations. It is possible that a GAN could create partially synthetic data and act as a more advanced way to ensure effective imputation. However, for our analyses, we are focused on developing fully synthetic PII.

B. Ethical Concerns in Modelling with PII

Bellovin et al. (2019) published a thorough review of synthetic data's legality and its strengths and limitations. As noted by Bellovin, one of the most robust methods used today to ensure the proper anonymization of data is through "differential privacy." Differential privacy offers a strict mathematical formula for computer scientists and researchers to maximize the accuracy of queries from a dataset while limiting or minimizing the potential for privacy leakage [5]. A core piece behind this formula is the introduction of randomness behind the imputation of certain features while not impacting the overall dataset's valid inferences. By introducing randomness to the data generation, differential privacy can create deniability behind identifiable features of individuals. The addition of deniability will be a core concept used in the approach we use to make our synthetic PII.

Bellovin et al. (2019) goes on to classify additional categories that define two types of synthetic datasets, one called "vanilla" synthetic data and the other called differentially private synthetic data. Data created from HealthGAN would fall into the bucket of "vanilla" synthetic data; data in, data out through a generative model. "Vanilla" synthetic data has the potential to be more at risk for over-inclusive privacy and under-inclusive privacy. The chance of identification may be improperly increased or decreased depending on the legal statute at hand, the methods used to train the model, and the ability to quantify the risk of identification. Either risk can affect the overall quality of the data or not meet privacy statutes, preventing any kind of public disclosure.

Differential privacy synthetic data can reduce the uncertainty "vanilla" synthetic data has [2]. From a legal standpoint, differential privacy synthetic data can assure that individuals will not be identified. And from a researcher's perspective, an analyst or researcher would have the assurance that the data remains useful. Predictive applications could still be trained on this synthetic dataset and used in real-world applications.

C. *A Novel Approach to the Generation of Synthetic Differential Privacy Data*

No research has shown a solution to generate synthetic PII data and promise minimal risk to privacy or anonymity. In this analysis, we propose a new and novel approach to generate synthetic differential privacy data. The algorithm developed is focused on a data context for the travel industry but could be easily adjusted for the healthcare industry or others as needed to be generalizable. We walkthrough how entropy is carefully handled in the algorithm to create the differential privacy attribute. Later, the data generated from the algorithm is discussed how it can be applied for training machine learning models or used for finding insights in data analytics.

III. Differential Privacy Fully Synthetic Data Generation

As mentioned in the introduction section, our aim was to generate a real-world travel dataset used to predict the PII features of an individual that may travel between Mexico and the United States. This predictive problem would likely require several PII-related features such as demographics, age, gender, dates of travel, and origin/destination locations. Additional features could also be added depending on the context of this problem by industry type. To ensure differential privacy in the fully synthetic dataset, we leveraged a Python library that could specialize in generating it, such as "Faker" or "Mimesis." Both packages can generate fake names of individuals and random addresses from a variety of countries. However, the "Mimesis" package has a higher likelihood of creating unique names. Out of the generation of 100,000 records, 98% were unique. Using "Faker," out of 100,000 records, only 71% were unique. For our dataset, we wanted the possibility of having multiple trips for the same individual; therefore, we used "Faker" as our core package in differential privacy fully synthetic data generation script. Table 1 depicts the structure of the dataset that is created.

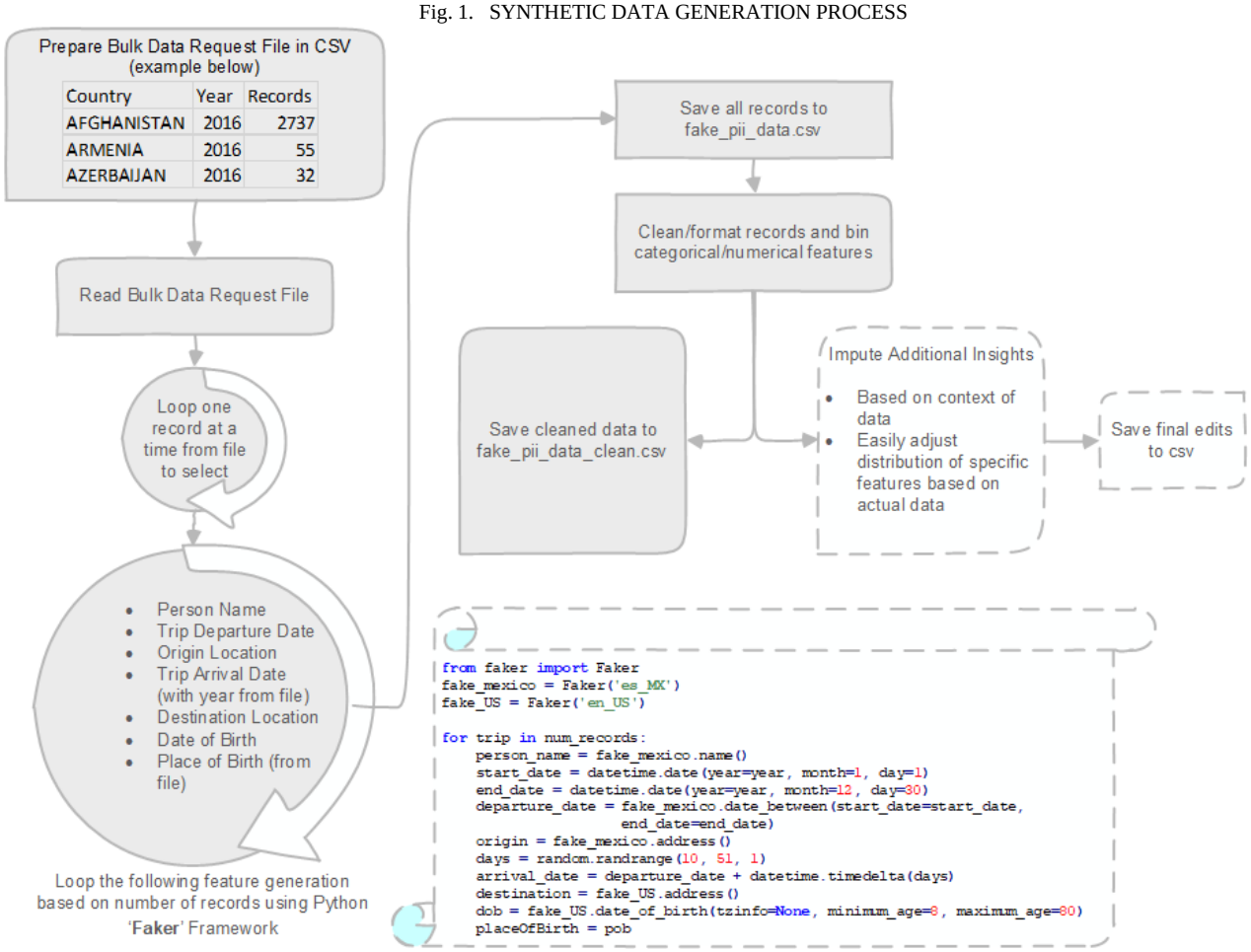
TABLE I. Synthetic Data Structure

Feature	***Data Type***	***Description***
person_name	String	Full name of a person
origin_id	Numerical	Random ID assigned to trip
departure_date	Datetime	Date of departure
origin_address	String	Mexico address from departure
origin_city	String	Mexico city from departure
origin_state_province	String	Mexico province from departure
origin_zip	Numerical	Mexico zip from departure
destination_id	Numerical	Random ID assigned to trip
arrival_date	Datetime	Date of arrival to the destination
destination_address	String	U.S. address of the destination
destination_city	String	U.S. city of destination
destination_state_province	String	U.S. state of destination
destination_zip	Numerical	U.S. zip of destination
trip_length	Numerical	Number of days of travel time
date_of_birth	Datetime	Date of birth of the person
age_at_arrival	Numerical	Age of person upon arrival
departure_year	Numerical	Year travel started
place_of_birth	String	Country of birth for person
gender	String	Gender of person
Season	String	Season during travel dates
AgeBin	String	Binned range of age of the person
TripLengthBin	String	Binned range of travel time
Feature	Data Type	Description
person_name	String	Full name of a person
origin_id	Numerical	Random ID assigned to trip
departure_date	Datetime	Date of departure
origin_address	String	Mexico address from departure
origin_city	String	Mexico city from departure
origin_state_province	String	Mexico province from departure
origin_zip	Numerical	Mexico zip from departure
destination_id	Numerical	Random ID assigned to trip
arrival_date	Datetime	Date of arrival to the destination
destination_address	String	U.S. address of the destination
destination_city	String	U.S. city of destination
destination_state_province	String	U.S. state of destination
destination_zip	Numerical	U.S. zip of destination
trip_length	Numerical	Number of days of travel time
date_of_birth	Datetime	Date of birth of the person
age_at_arrival	Numerical	Age of person upon arrival
departure_year	Numerical	Year travel started
place_of_birth	String	Country of birth for person
gender	String	Gender of person
Season	String	Season during travel dates
AgeBin	String	Binned range of age of the person
TripLengthBin	String	Binned range of travel time

The Python library "Faker" created the features: person name, date of birth for person, departure and arrival dates, origin location, and destination location. A total of 5.665% of the records were duplicated person names. Each origin location is an actual location in Mexico, and each destination location is a real location in the United States. The remaining features were calculated and cleaned. The gender feature was calculated using another Python library called "Gender Guesser," which can categorize a string name of a person and assign either female, mostly female, male, mostly male, or unknown. We later adjusted the results to female from mostly female and male from mostly male. Any unknown assignments were fixed to male due to the context distribution of predominantly males being the population of migrants traveling between Mexico and the United States [8].

To make the distribution of travel records further represent real-world context, the distribution for the place of birth for all individuals is based on the public dataset "Refugees and Asylees 2019 Data Tables" provided by the U.S. federal government agency, Immigration Customs Enforcement [19]. The distribution of data provided by ICE extends from 2010 to 2019 by country. However, to decrease the performance time to generate the data, we only used the distribution data from 2016 to 2019, totaling 190,649 individual trips.

Fig. 1. SYNTHETIC DATA GENERATION PROCESS



Additionally, all ages of individuals were later adjusted by "Faker" to reflect the distribution of ages from the data provided by ICE, then binned into the following buckets: 0 to 9, 10 to 19, 20 to 39, 40 to 60, and 60 years and over. Figure 1 above, depicts a diagram of the algorithm's process to create the fully synthetic PII data.

The algorithm developed will read a CSV file that includes country, year, and number of records. Then for each country and year, generate the number of synthetic records noted in the file. One of the minor flaws with using the Python library "Faker" is that some of the U.S. destination addresses created were P.O. Boxes or Military addresses that do not reflect city or state sometimes. The synthetic records created with those destinations (20,408) were eventually dropped from the dataset. Therefore, the total distribution of individuals travelled is only 89% accurate to the actual distribution in ICE's public data table. Our 170k records took approximately 8hrs to generate. Table 2 below depicts the length of time for data generation for 1000 records, 10,000 records and 100,000 records.

TABLE II. TIME TO GENERATE DATA FROM ALGORITHM

Generation Time	**Number of Records**		
	1,000 Records	***10,000 Records***	***100,000 Records***
Time Used	3mins 31 seconds	35mins 33seconds	5hrs 38mins 24 seconds

Since all 170,241 records were synthetically generated with randomized PII values, using the Python libraries of "Faker" or "Mimesis" might be the answer for creating fully synthetic differential privacy data. These tools can be used broadly in most industries and can even provide additional PII information such as e-mails, phone numbers, social security numbers, and even credit card numbers. For our focused analysis, these features would not be necessary. In addition, the creation of this fully synthetic dataset only required knowledge of the structure needed based on context. No complex generative models were required to be trained from actual data to create new data. This process could be valuable for creating practice datasets for analyses or predictive modelling from an academic perspective.

In the next section, we prove that this fully synthetic differential privacy dataset assures valid inference to analysts and some of the potential insights that could be made by organizations interested in predicting the PII features of individuals. We also demonstrate that the quality of the synthetic data generated can still be useful for predictive modelling.

IV. VALIDATING STATISTICAL INFERENCE

Knowing whether two variables are correlated and thus substitutable is helpful for understanding data variance structures and feature selection in machine learning. Also, it is essential to understand the associations between variables for data analysis and hypothesis testing. Since most of the PII fields in the fully synthetic dataset are categorical or binned into category groups, a first step in identifying insight is to understand any possible correlation between those categorical features. Table 3 depicts a breakdown of the feature types of which ones were purely randomly generated and which ones were random but based on the distribution of a real dataset.

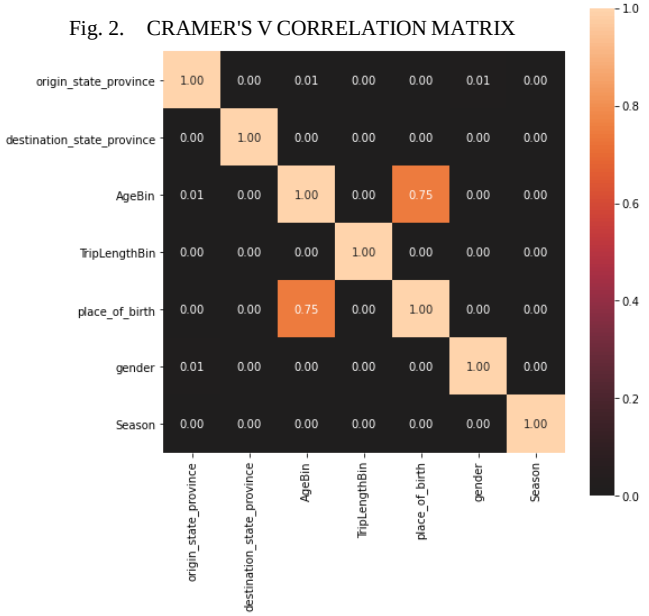
TABLE III. DATA FEATURE DISTRIBUTION METHOD

Feature	*Type*	*Distribution Method*
origin_state_province	Categorical	Random
origin_state_province	Categorical	Random
AgeBin	Categorical	Real Distribution
TripLengthBin	Categorical	Random
place_of_birth	Categorical	Real Distribution
gender	Categorical	Real Distribution
season	Categorical	Random
trip_length	Numeric	Random
age_at_arrival	Numeric	Real Distribution

When comparing two categorical variables, we can quickly transform the original vectors into contingency tables by counting the frequencies of the categories. Contingency tables are widely used in scientific research across disciplines to represent the multivariate frequency distribution of variables. Because of their widespread use in statistical studies, a family of tests has been developed to assess the significance of categorical differences.

To calculate the correlation/strength-of-association of features in the dataset, we select Cramer's V and Theil's U for nominal data with no intrinsic order. Cramer's V is based on a negligible variation of Pearson's Chi-Square Test and comes with some benefits. The output of Cramer's V is in the range of [0, 1], where 0 means no association and 1 means complete association. However, it can reach 1 only when each variable is wholly determined by the other. Unlike correlation, there are no negative values, as there is no such thing as a negative association. Like correlation, Cramer's V is symmetrical, which means it is insensitive to swapping x and y. Fig 2 depicts the results from a Cramer's V statistical test.

Fig. 2. CRAMER'S V CORRELATION MATRIX

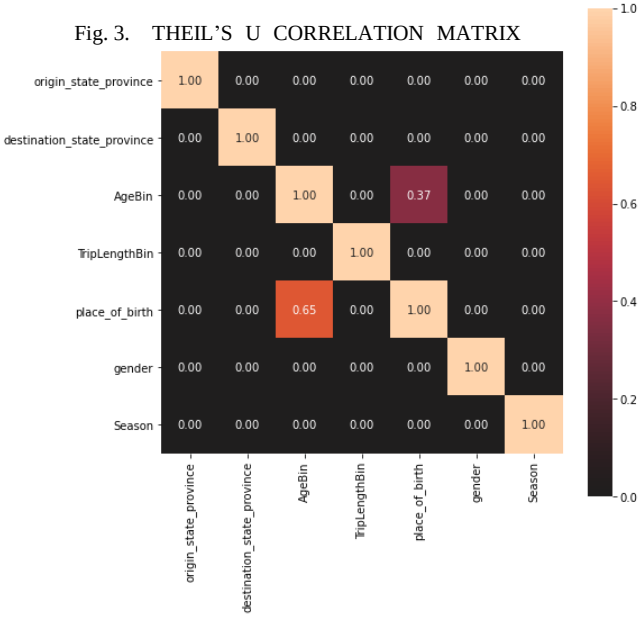


Besides Cramer's V, we also use the Theil's U test as an asymmetric measure of association between categorical features. Theil's U, also known as the Uncertainty Coefficient, is based on the conditional entropy between x and y. It is a measure of association that indicates the proportional reduction in error when values of one variable are used to predict the value of the other variable. Like Cramer's V, the output value is in the range of [0, 1] with the same interpretation. However, unlike Cramer's V, it is asymmetric, meaning

$$U\,(x,y) \neq U\,(y,x) \quad (1)$$

(while $V\,(x,y) = V\,(y,x)$, where V is Cramer's V). Fig 3 depicts the results from a Theil's U statistical test.

Fig. 3. THEIL'S U CORRELATION MATRIX



There is a significant association between birthplace and age group from two correlation tests. In Cramer's V, the strength of association is 0.75, indicating a strong association between two features. And from Theil's U output, the value of 0.65 suggests that knowledge of birthplace reduces error in predicting age group by 65%. And it reduces by 37% another way around.

From the data generation perspective, the strong association can lead to biased interpretation. The distribution of age bin group and birthplace are entirely independent. In the migrant travel case, no information shows that people of a particular age group are more likely from one birthplace. However, even using the fully synthetic data, we could easily associate PII fields and make a biased conclusion.

In the next section, we discuss the results of an unsupervised machine learning concept called association rule mining that can be used to predict the travel destination individuals may take based on their PII or other feature associations. In addition, we state potential risks that could occur if additional features were added for predictive modelling using a neural network.

V. UNSUPERVISED MODELLING WITH PII

Association rule learning is a rule-based machine learning method for discovering interesting relations between features in large databases. It identifies frequent if-then associations called association rules, consisting of an antecedent (if) and a consequent (then). The unsupervised modeling concepts also aim to show how frequently an item set occurs in a transaction. The objective of our modeling is to see whether we can get any association between travel origins, destinations, and PII fields.

The first step was to implement one-hot encoding on all the categorical features in the synthetic data to prepare for association rule mining. A one-hot encoding is a representation of categorical variables as binary vectors. The

basic strategy for one-hot encoding is to convert each category value into a new column and assign a 1 or 0 (True/False) value. This benefits from not weighing a value improperly but does have the downside of adding more columns to the data set.

In the modeling, we apply the Apriori algorithm for extracting frequent item sets for further analysis. The Apriori algorithm uses frequent item sets to generate association rules, and it is designed to work with transaction data. With the help of these association rules, we can determine how strongly or how weakly two objects are connected. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset associations efficiently by reducing the search space. There are three critical measurements in associate rule learning: support, confidence, and lift. Support is the relative frequency that the rules show up. High support ensures the generalization of the relationship. However, there may be instances where low support is helpful to uncover some "hidden" relationships. Confidence is a measure of the reliability of the rule, and it is the conditional probability of occurrence of the consequent given the antecedent. Confidence of 0.8 means that in 80% of the cases where antecedents happen, consequents will also occur. The lift metric is commonly used to measure how often the antecedent and consequent of a rule occur together than expected if they were statistically independent. The basic rule of thumb is that a lift value close to one means the antecedent and consequent were completely independent; thus, lift values greater than one give a more valuable rule pattern.

Considering the uneven distribution within some features, we set 1% as the support threshold in our synthetic data. Then we output the rules with their corresponding support, confidence, and lift. We have 64 rules with a confidence of more than 90%. Table 4 depicts some examples from the output.

TABLE IV. APRIORI ALGORITHM AGE RESULTS

antecedents	*consequents*	*support*	*confidence*	*lift*
place_of_birth _AFGHANISTAN	AgeBin_0-9	0.0219	1	3.7700
place_of_birth _BHUTAN	AgeBin_0-9	0.0615	1	3.7700

The two rows from Table 4 above show 100% confidence and a high lift value. The migrant travel case says that if the individual's birthplace is from Afghanistan or Bhutan, he/she has a 100% of chance in the 0 to 9 age group when travelling from Mexico to the United States. There are a lot of rules related to birthplace and age group with high reliability from the model. On the other hand, it also explains why those two features have a high association in the statistical test discussed earlier.

On a matter of travel origin and destination, which are generated randomly, we also have some interesting rules from the model. Table 5 below depicts examples of origins/destinations likely to have a traveler by a specific gender. The conclusions are not consistent with how the synthetic data was generated since the distribution of travel records originated from a real dataset. The conclusions in Table 5 are new and unique from the data generation process.

TABLE V. APRIORI ALGORITHM GENDER RESULTS

antecedents	*consequents*	*support*	*confidence*	*lift*
origin_state_province _GRO	gender _male	0.0193	0.6222	1.0544
destination_state_province _MD	gender _male	0.0118	0.6151	1.0349
origin_state_province _BCS	gender _female	0.0130	0.4174	1.0479
origin_state_province _CHIS	gender _female	0.0127	0.4104	1.0303

The reason behind the misleading result is that we do not have a balanced distribution of categories in features. From the contingency tables analysis, if we have few records on one birthplace, it is likely that age groups are not randomly aligned within that birthplace. It could easily lead us to the biased conclusion that a traveller born in one place belongs to one age group. The unsupervised model, such as association rule mining, can make the biased judgment more specific since the model is only based on the data distribution. Even with entirely random objectives (origin/destination), a model still remains a risk to reveal biased insights.

VI. CONCLUSION

This analysis can conclude that creating fully synthetic, differential privacy PII data may be possible using existing programmatic frameworks like "Faker." However, additional post-processing must be completed to ensure that the randomized features are similar to distributions of real data in order to act as a true replica. This process can easily be customizable based on the context of categorical data needed. It may prove helpful for added data in training machine-learning models or practicing analytics in an academic environment.

In addition, we can conclude that creating unbiased associations or predictions when using PII data for models may be difficult to avoid. With PII features generated from the real distribution, the correlation test could show some strong associations due to frequencies of occurrence. Those associations cannot be generalized for a new batch of generated synthetic data since we imputed additional insights for the context of our data. However, it could bring potential risks when performing the exploratory data analysis and predictive modelling on the synthetic PII data. The insights that we learned from synthetic data and adjusted distributions of real data might lead to some coincidental findings that could mimic genuine insights of real data. The process we took of adding additional insights is the generalizable concept that can be applied to a multitude of topics for academia or industry.

VII. DISCUSSION

As data continues to grow, industries and government institutions may want to leverage PII data more for predictive applications. However, it is improbable for most organizations to discuss these issues or their modelling results in a public forum using this type of data due to ethical/privacy concerns. There may be more of a need for the academic community to partner with privacy protection institutions to ensure enough regulation and caution within organizations using PII data for predictive purposes.

Just recently, we know U.S. law enforcement agencies are targeting social media companies like Instagram, Facebook, and Twitter more than usual for user's data [3]. It's entirely possible law enforcement agencies are interested in leveraging PII more than ever, using machine learning to predict the

where, when and who in crime, like Steven Spielberg's famous Sci-fi/Action movie *Minority Report*. While it may be in the best interest of public safety, avoiding biased associations or predictions may be a significant risk entirely ignored by these organization's data scientists and analysts.

Using the context of our analysis, private or public organizations in the business of travel may collect other PII features of users, such as their interactions on platforms in booking travel. How often someone views airfare, hotels, or train reservations during a particular timeframe may indicate their future interests. This type of information tied to their personal profile on a booking website containing travel history and then adding social media data scraped from the internet may create a very targeted advertisement to individuals. This kind of prediction, primarily fueled by PII, can conflict with the regulations that protect individual privacy.

As the next step from this analysis, additional research should evaluate our proposed fully synthetic, differential privacy data generation process. Adding other PII features may create additional complexity to how inference is made. Exploring embedding as an additional step for privacy-preserving may prove helpful [16]. In addition, further review and analysis should continue with legal researchers and data scientists to propose potential solutions to how PII data should be handled in the development of predictive applications by private and public institutions.

All code and data made for this publication can be found at: https://github.com/mwilchek/Differential-Privacy-Synthetic-PII.

ACKNOWLEDGMENT

This research and analysis were supported by the guidance and review of the Data Science & Analytic program's professors at the George Washington University, George Mason University and Georgetown University. We thank Dr. Jacqueline Serigos (George Mason University) and Dr. Amir Jafari (George Washington University) for reviewing this paper and providing helpful feedback.

REFERENCES

[1] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, MenglingFeng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo AnthonyCeli, and Roger G Mark. Mimic-iii. (2016) A freely accessible critical care database. In *Scientific data, 3:160035*.

[2] Bellovin, S. M., Dutta, P. K., & Reitinger, N. (2019). Privacy and synthetic datasets. In *Stan. Tech. L. Rev., 22*, 1. https://law.stanford.edu/wp-content/uploads/2019/01/Bellovin_20190129.pdf

[3] Bhuiyan, Johana. (2021, March 24). This is what happens when ICE asks Google for your user information. *Yahoo News*. https://news.yahoo.com/happens-ice-asks-google-user-120018014.html

[4] Fienberg, S. E., Makov, U. E. and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *J. Off. Statist. 14, 485-502*. https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/disclosure-limitation-using-perturbation-and-related-methods-for-categorical-data.pdf

[5] Hilton, M. (2002). Differential privacy: a historical survey. In *Cal Poly State University*. https://perma.cc/J3HT-DMWB DOI:https://doi.org/10.1145/1013115.1013129

[6] Hong, Jason I., Jennifer D. Ng, Scott Lederer, and James A. Landay. (2004). Privacy risk models for designing privacy-sensitive ubiquitous computing systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques* (*DIS '04*).

[7] Little, R. J. A. (1993). Statistical analysis of masked data. In *J. Off. Statist.* https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf

[8] Montes, Juan and Alicia A. Caldwell. (2021, March 24). Men Looking for Work Drive Surge in Illegal Crossings at the U.S. Border. *The Wall Street Journal*. https://www.wsj.com/articles/men-looking-for-work-drive-migrant-surge-at-the-u-s-border-11616624482?reflink=desktopwebshare_permalink

[9] Onik, Md Mehedi Hassan & Al-Zaben, Nasr & Yang, Jinhong & Lee, Nam-Yong & Kim, Chul-Soo. (2018). Risk Identification of Personally Identifiable Information from Collective Mobile App Data. 71-76. In *International Conference on Computing, Electronics & Communications Engineering (iCCECE '18)* 10.1109/iCCECOME.2018.8659213

[10] Plummer, Libby. (2017, August 22). This is how Netflix's top-secret recommendation system works. In *Wired*. https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like

[11] Qiu, Han., Qiu, Meikang., Lu, Zhihui. (2020). Selective encryption on ECG data in body sensor network based on supervised machine learning. In *Information Fusion 55, 59-67*. http://paper.idea.edu.cn/paper/2965302537

[12] Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Statist. 19, 1-16.* http://www2.stat.duke.edu/~jerry/Papers/jos03.pdf

[13] Reiter, J., & Drechsler, J. (2010). Releasing Multiply-imputed Synthetic Data Generated In Two Stages To Protect Confidentiality. In *Statistica Sinica, 20*(1), 405-421. http://www.jstor.org/stable/24308998

[14] Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *J. Off. Statist.* https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

[15] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3-18). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7958568

[16] Xiang, Xiayu., Duan, Shaoming., Pan, Hezhong., Han, Peiyi., Cao, Jiahao., and Liu, Chuanyi. (2020). From One-hot Encoding to Privacy-preserving Synthetic Electronic Health Records Embedding. In *Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies* (*CIAT 2020*). DOI:https://doi.org/10.1145/3444370.3444605

[17] Yale, Andrew., Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. (2019). Assessing privacy and quality of synthetic health data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse* (*AIDR '19*). DOI:https://doi.org/10.1145/3359115.3359124

[18] Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. (2019, April). Privacy preserving synthetic health data. In *ESANN 2019-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. https://hal.inria.fr/hal-02160496/file/ESANN_2019.pdf

[19] Zhang, Zhiguo., Wu, Jingqi., Deng, Jing., Qiu, Meikang. (2008). Jamming ACK attack to wireless networks and a mitigation approach. In *IEEE GLOBECOM pp 1-5*. https://www.uncg.edu/cmp/faculty/j_deng/papers/jack_globecom08.pdf

[20] Immigration Customs Enforcement (ICE). (2020). Refugees and Asylees 2019 Data Tables. Immigration Statistics. https://www.dhs.gov/immigration-statistics/refugees-asylees