Anise Mulkey, Zach Brown, Kumar Raja Guvindan, Cory Nichols

[[1]](#footnote-1)

A Data Tug of War: Privacy and Data Analysis

***Abstract* – Individual privacy is both an ethical consideration and a personal expectation. As a logical extension, the privacy of an individual’s personal *data* is both an ethical consideration and personal expectation. However, the importance of data analysis and the desire for insights from data has exponentially increased the need for access to and storage of massive amounts of data. Much of this data is stored on the Internet and made publicly accessible. The proliferation of data made publicly available today is a threat to individual privacy.**

**Personally Identifiable Information (PII) can be acquired without authentication and used to identify individuals’ health or financial status, interests or even more private information such as a social security number. PII is often valuable in the analysis setting and presents a supposed opportunity for better analysis results.**

**In this paper, we explore the privacy versus analysis trade off. Specifically, we explore the impact anonymizing financial data has on the results of a classification model. We investigate the anonymity of a publicly available dataset of loans and compare the effectiveness of a data analysis method on the original loans dataset and three anonymized datasets. For the described classification task, two out of three anonymization techniques result in no loss of analysis integrity, providing evidence that in this scenario, certain data anonymization techniques do not affect the results of data analysis.**

I. INTRODUCTION

P

ERSONALLY identifiable information (PII) is both an asset and a liability. It can be part of a contract, taken illegally, and even given away for free, knowingly or unknowingly. PII is a treasure for enterprises, identity thieves and hackers alike. It can be

used by itself or in combination with other information to identify an individual. Given enough data, such as a work title, address and age, or a unique identifier such as a social security number, a person’s identity can be discovered and used for marketing or more sinister purposes.

Marketers can tailor campaigns and product offerings based on a person’s income level, zip code, gender, preferences or a combination of features often available in public datasets. Given personal data for a patient, medical information can easily be mined to identify a person and their associated medical history. If a dataset lacks information, additional features can be appended to the original dataset by using public datasets or performing statistical analysis methods to personally identify individuals [1]. An analyst does not need to work in a confidentially-privileged role to obtain personally identifiable information. Public datasets used for analysis are readily available. These datasets present dangerous possibilities for identifying the people associated with records, threatening individual privacy and presenting serious ethical considerations for the analyst.

Healthcare research often releases supposed anonymized datasets used for the verification of research results. The goal of anonymization is to protect the identity of individuals used in research. However, combining these anonymized research datasets with publicly available datasets such as tax records, census data or public membership lists has been shown to increase the possibility of identifying an individual in the anonymized research data set [2].

A consumer driven example of unwanted personal identification involves credit card metadata. Analysts were able to accurately re-identify credit card users in shopping malls using nothing other than their purchase histories, location, date and price of each transaction. Unique identifiers were used to label each customer. The researchers in this case did not need to use other data sources outside of the transactions and transaction metadata [3].

Individual privacy is an expectation; privacy must be protected at all times, even when data analysis is the main objective. Individual privacy is protected when the data is considered anonymous. This means no relationship can be established between a person’s identity, such as their full name, and the data available. However, the very nature of data analysis places critical importance on accurate results. By truly anonymizing a dataset, the analyst risks losing critical information for data analysis. Therefore, a tradeoff exists between privacy of the data and accuracy of results when analyzing personally sensitive datasets.

In this paper, we de-anonymize records in a public dataset of personal loans by appending data available from a public social media API. We analyze the original loans dataset and transform features with high re-identification risk using anonymization techniques to produce three private datasets. To address the previously mentioned privacy-analysis tradeoff, we perform a classification task using both the anonymized loan datasets and the original loan dataset to predict the binary outcome of a loan. The results from each classification task are examined to determine if anonymization has an impact on analysis results. Our objective is to determine whether making the loans dataset safer impacts the accuracy of classifying the outcome of a loan.

The dataset used in our analysis is a publically available dataset of personal loans retrieved from Kaggle.com. The dataset contains 887,379 unique loans, with a mixture of 74 categorical and quantitative features. These features include the loan’s interest rate, a person’s home ownership status, employment length, job title, annual income and the outcome of the loan.

Given an analysis task of classifying the outcome of a loan, we are able to reduce re-identification risk, increase privacy and maintain analysis accuracy results for two of three anonymization methods. However, anonymization in the two successful scenarios is not completely successful, leaving significant re-identification risk in the resulting datasets. While analysis was not impacted in these scenarios, additional anonymization methods may need to be explored to truly anonymize the loans dataset.

Our paper is organized as follows: In Section II, we give a brief overview of related work and describe various anonymization techniques and frameworks available for PII data today. In Section III, we provide a detailed overview of our research and analysis methodology. In Section IV, we prove out the risk associated with the loans dataset, perform exploratory data analysis, implement three anonymization techniques and fit logistic regression models: one based on the original dataset and three based modified loans datasets with key PII features controlled. Ultimately, we compare these models to interpret the impact anonymizing key PII features has on data analysis. In Section V, we summarize our findings and address the limitations of our work.

II. RELATED WORK

A common example of data privacy standards and protection of PII comes from the Department of Health and Human Services (DHHS) in the United States. DHHS standardized the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule to protect personal health information. The Privacy Rule’s goal is to “assure individual’s health information is properly protected while allowing the flow of health information needed to provide health care.” It protects the privacy of people who seek and receive care. In particular, HIPAA protects “individually identifiable health information” and applies data anonymity standards to demographic information relating to past, present and future physical or mental health, the provision of health care and financial information. The Privacy Rule also sets standards for health data disclosure, authorization of usage and establishes severe penalties for violation of the Privacy Rule [3].

HIPAA has been established as a baseline for protecting data in industries outside of healthcare, such as the financial industry. Under frameworks such as HIPAA’s privacy rule, brute force methods such as data truncation or outright data exclusion have been utilized in the past to achieve anonymity of personally identifiable information in datasets containing people as objects or rows. Telephone numbers, addresses, social security numbers and credit card information are deleted from the dataset or truncated to a number of digits that represent an abstraction of the original data [5]. However, the effects of brute force anonymization techniques, especially the deletion of features, can severely impact the results of data analysis if care is not taken to maintain the statistical relationships of features in a dataset.

Related work has been completed to address the analysis-privacy trade off. In a paper from the Office for National Statistics in the United Kingdom [6], administrative data from a national healthcare source was anonymized such that personal privacy was maintained but data captured from the health care source could be re-used for census analysis purposes. The goals of the research team were to find a methodology to provide more frequent census analysis capabilities while cutting costs of clerical matching and ensuring citizen privacy. A cryptographic hash function was used to anonymize personally identifiable information retrieved from administrative datasets such as the National Health Register. These anonymous data were linked with existing census data to build a statistical representation of various counties in England and Wales. The research team used two methodologies: logistic regression and rules based matching on the anonymized dataset to confirm a match with census data, while protecting the true identities and health status information about each person in the dataset. Researchers measured the quality loss of their methodology by comparing against “gold standard,” clerical methodologies. The anonymous matching model performed only slightly worse in estimating actual census data for eight local authorities tested.

A well-known example of combining data to identify users involves data used in the Netflix challenge. The Netflix Challenge was a competition held to provide Netflix with a better recommendation system for movies. While the data used for the competition was considered anonymous by Netflix, user names were stripped out, analysts were able to combine internet movie database (IMDB) data, including user names and profile information with the “anonymous” Netflix data. This allowed the analysts to identify many of the users in the Netflix dataset. This information allowed the researchers to determine a user’s political orientation, religious views and even body type based on movie reviews [6].

In our case, PII associated with an individual’s finances is a major privacy concern. Financial privacy s is a baseline consumer expectation. Privileged users may have access to financial PII as part of their jobs or a contract. However, in the public space, it is problematic if user identities can be mined from publicly available financial datasets. Therefore, public personal loans data should be protected, even if the data is to be used for analysis or research. An analyst, hacker or identity thief could easily violate privacy expectations of consumers and make an assessment on a person’s credit-worthiness, or worse, impersonate the person on other financial applications for loans if given enough data. In the realm of data analysis, however, additional data is an asset. Data privacy may actually impede analysis, creating the tradeoff previously mentioned between privacy and meaningful analysis results.

III. ANALYSIS METHODS

Our objective is to determine whether making the loans dataset anonymous impacts the accuracy of classifying the outcome of a loan. The methodology and analysis for our research can be divided into three key tasks: 1) risk assessment and anonymization of personally identifiable features, 2) pre-processing of the loans dataset and 3) the use and performance assessment of a supervised machine learning technique to predict loan outcome on the original and anonymized loan datasets. For the anonymization task, we empirically investigate the risk of the loans dataset, re-identifying individuals using a social media API. We also utilize an open source tool, ARX, to gauge the risk of features in the loans dataset and subsequently transform PII features. In the data pre-processing phase, we address missing values in our data, delineate strategies for handling categorical data, extract the most meaningful features and eliminate redundancy, conduct feature scaling where appropriate and partition our data into training and test sets. Finally, we fit logistic regression models to classify the outcome of the loan and interpret the impact anonymization has on data analysis.

1. *Re-Identification Risk Assessment*

To assess and mitigate the re-identification risk associated with the loans dataset, two steps are taken. First, an empirical approach is utilized to assess the identification risk of the loans dataset by appending social media data to the original dataset. This approach is a simple, straightforward method for proving re-identification risk associated with the loans dataset. We utilize Facebook’s Graph API and a linear rules-based matching algorithm to marry social media data with loan records. A sample of 75 loans is taken from loans with creation dates in 2015 or later and ran through the matching algorithm to couple users to loan records based on a confidence rating.

Second, data re-identification risk is directly assessed and anonymization methods are used to create three secondary, anonymized datasets. To determine re-identification risk and subsequently anonymize the datasets, we use ARX [8], open source software for measuring re-identification risk and anonymizing PII data.

To assess data re-identification risk, ARX utilizes two different models: the prosecutor and the journalist model. In the prosecutor model, an intruder (prosecutor) knows that a particular individual (defendant) exists in an anonymized database and wishes to find out which record belongs to that particular individual. In the journalist model, the attacker can re-identify an arbitrary individual. The intruder does not care which individual is being re-identified, but is only interested in being able to claim that it can be done. The journalist scenario is similar to the social media task undertaken earlier in this section. These two models are ultimately used to assess re-identification risk based on quasi-identifiers, or combinations of features resulting in unique records. ARX also uses 18 HIPAA identifiers, including names, zip codes, IP addresses and social security numbers to assess re-identification risk. Quasi-identifier, HIPAA and other descriptive measures are combined to provide an overall risk for re-identification at the record level as well as for the entire dataset for each model type.

1. *Anonymization Techniques Employed*

We apply three different anonymization algorithms to the loans dataset via ARX: K-Map, population uniqueness and sample uniqueness. Each algorithm produces a singular dataset, resulting in three unique datasets to be used in the classification task for loan outcome.

K-Map is a specific implementation of K-anonymity. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release [18]. For example, if k = 5 and the potentially identifying variables are age and gender, then a k-anonymized data set has at least 5 records for each value combination of age and gender.

Based on a study by L. Sweeney [11], 87% of the US population can be uniquely identified by gender, ZIP code and full date of birth population (based on 1990 census data). For the population uniqueness method re-identification risk is estimated statistically by comparing the threat of disclosing various demographic data to a database of census data and the ability to re-identify in based on each scenario. For example date of birth, assuming that births are uniformly distributed across the days of the year. With this assumption, if n people are born in a given year, the expected number of days on which k individuals are born is given by [17]:

(1)

In the sample uniqueness method, the current dataset itself is used as the reference. All three methods, are used as metrics to assess the “attacker models” mentioned in the *Re-Identification Risk Assessment* section.

*Pre-processing and Exploratory Data Analysis*

After addressing the re-identification risk for the entire dataset, we prepare for our classification task by pre-processing the data. First, we filter our dataset to represent unsecured loans originated during a five-year period from 2007 to 2011. This particular time period is representative of wide variations in economic cycles, including the housing bubble burst of mid-2007, the sub-prime mortgage implosion and market correction and the ensuing recession and recovery. The macro-economic environment is a confounding variable in any loan performance evaluation and selection of origination vintages across this period of eight years reduces the impact of the macro-economic swings seen each year.

Examination of the data reveals that all 74 attributes in the original dataset are not necessary for classifying loan outcome. Several features are null or duplicated, and some of the data is not meaningful to classification of loan outcome, such as the loan identification number. These features were removed entirely from the dataset. We also perform the following pre-processing and imputation steps to convert critical feature columns into consumable data for our classification task:

* Changed loan term to continuous and deleted the string 'mos'
* Changed employment length to numeric and imputed 5062 values less than 1 year to 1 year
* Changed earliest credit line to continuous number of months since earliest credit line, using today’s date minus earliest credit line date

Additional data imputation is performed on features such as the loan’s credit line amount and employment length. Mean value imputation is applied in these cases to fill null values for these quantitative features with their respective mean values across the data set. Correlated features such as sub-grade and overall grade of the loan are compared and stronger linear correlation scores with loan outcome determine which features to keep for analysis. Categorical features such as loan purpose, home ownership and income verification status are converted into Boolean features using one hot encoding.

One feature of particular importance is annual income. This feature has a material number of outliers and an obvious right skew. To proactively address the issues this will introduce into our analysis, we log transform annual income and present the outcome of that modification below. The resulting distribution is much more appropriate for linear analysis:

*Figure 2.1: Distribution of Log Transformed Annual Income*



Finally, we change loan status to a binary loan disposition indicator subject to conditions in the table below:

TABLE I

Loan Status Result: Binary Transformation

|  |  |
| --- | --- |
| **Indicator** | **Original Loan Status Description** |
| Bad | Charged Off |
| Good | Current |
| Bad | Default |
| Bad | Does not meet the credit policy.  Status: Charged Off |
| Good | Does not meet the credit policy.  Status: Fully Paid |
| Good | Fully Paid |
| Good | In Grace Period |
| Good | Late (16-30 days) |
| Bad | Late (31-120 days) |

In total, 23 features and 42,535 records are kept after exploratory data analysis. Our truncated feature set represents the most viable data points for data analysis purposes. However, it also maintains critical PII features such as employee title, work experience, and zip code [9].

1. *Machine Learning Method Utilized*

After data pre-processing, logistic regression is utilized to perform loan outcome classification. Specifically, we are interested in whether the loan result was “good” or “bad” (late, default) as defined by table 1. Logistic regression is a linear classifier utilizing coefficients or weights for features and a link function called the sigmoid to transform continuous responses based on those weights to probabilities for each loan outcome class. Using fitted weights and feature values, we predict the previously mentioned binary categorical feature, loan outcome, for both the anonymized loans datasets and the original dataset. We measure the success of prediction using accuracy and F1 metrics, which gauge how well our model predicts overall and how likely our model is to predict the correct outcome, respectively.

The loan outcome is split unevenly, with 85% of records resulting in a “good” outcome and 15% of records resulting in a “bad” outcome. To address this tendency for overfitting and optimize model fits across raw and anonymous datasets, we fine-tune the performance of our models by adjusting regularization parameters and a penalty strength parameter.

IV. ANALYSIS RESULTS

1. *Lack of Data Anonymity*

An empirical approach is used to prove the re-identification risk of the loans dataset. Some of the features are transformed to protect loan applicants’ privacy. However, these transformations are not adequate and maintain risk in the dataset. We utilize Facebook’s Graph API [10] to test the anonymity of a sample of the original loans data using an algorithmic approach. Of 75 loan applicants sampled, nine individuals are uniquely identified. This method does not require in-depth statistical analysis, rather, feature scoring is used to estimate the confidence in a match between the loans dataset and Facebook profile data.

Five key features are used in our query to match loans to Facebook user profiles: Age range in years, location, job title, work experience in years and loan create date. All of these features are available in our final dataset for analysis. Other viable PII data that can be used in chaining data sources together available in the loans dataset is income and home ownership status. These features were not used in our Facebook queries, as income and home ownership statuses are not standard fields available in user profiles. However, they would prove valuable for chaining other data sources together, such as property tax and census data [11].

Zip code in the loans dataset is partially hidden for privacy reasons. The first three numbers, however, are available and can be used to identify user city, state combinations applying for loans. We utilize publicly available zip code data [12] to capture a loan applicant’s city via the first three numbers in the zip code and use the state already available in the loans dataset to complete a full city, state location.

Further, job title is particularly risky, especially for unique job position titles at local businesses. Work experience allows us to infer the age of the user, or at least a range for age. Data such as location, recent activity and usage rates have been used previously to identify users by coupling Twitter data [12, 16].

These fields can be searched algorithmically with Facebook’s Graph API and a rules-based classification approach. API search queries in Facebook have been diminished over the past two years due to privacy concerns. A robust SQL-like interface is now removed in favor of a stripped-down wildcard search method with more limited capability. However, even with diminished search capabilities, a wildcard search is still capable of producing a list of user unique IDs. These unique IDs are then query-able and return fields we wish to match to the loans data to identify the owner of a loan. Users are rated as potential matches based on a simple linear weighted scoring method such as:

.

(2)

Match Score has a range of zero to five. Because of the specificity job title provides, a higher weight of two is used in the linear rules-based model. A job title match is attempted from the loans data set to a list of employment history from the user profile on Facebook. If the entirety of the job title is contained in employment history, the match is considered 100%, thus the JobTitle variable is scored as 2 in the equation. Partial matches are also allowed based on counts of matching keywords in a bag of words for each record. Location is binary and is matched based on city, state in the loans dataset and the Facebook profile.

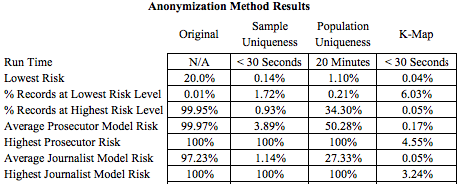
Outlining the lack of security in the loans dataset further, a loan applicant’s age can be inferred from the work experience listed on each loan application. An individual with eight years of work experience, for example, has an age range of 24 to 28. If a user falls into this category, it is scored highly (100% match). Secondary ranges are also used, with lower scores, to allow for fringe matches. For instance, the age range 23 to 29 is a secondary criterion, if matched, would result in a 75% match, adjusting the overall Match Score downward but not to zero. Similar logic is also applied to loan creation date in order to take into account the recency of the loan.

75 recent loans (two years or younger) were randomly sampled and scored utilizing the previously mentioned algorithm. We are able to uniquely match 9 Facebook profiles with user names and minimize an additional 17 loans to five users or less. Additionally, data including user group membership, personal interests and other personalized information from Facebook can be appended to the loans dataset. While accuracy of classification is not our objective in this section, we show that the loans dataset, a public dataset, presents significant re-identification risk. Facebook has deprecated unique ID look ups as part of the Graph API as of October 30th, 2016 and further limited the ability to capture user identities via their API. However, even a simple search on Facebook’s UI would allow user identification. We believe this still provides substantial risk when coupled with other PII data such as personal loans.

1. *Anonymization Techniques*

TABLE II

Re-identification risk by 3 anonymization methods

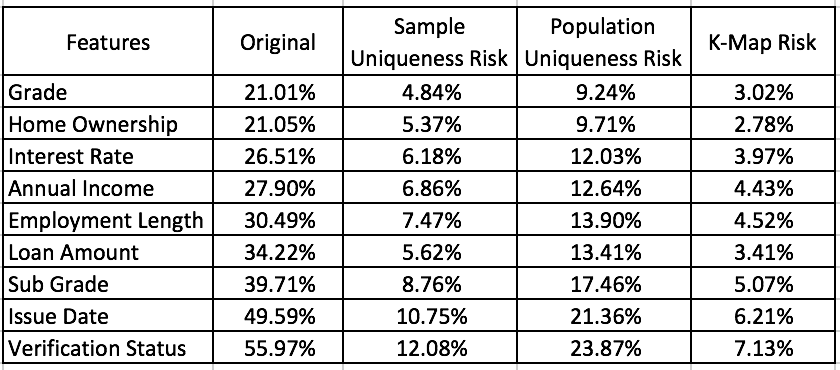


In order to make an objective assessment of the best method to reduce re-identification risk, a simple scorecard is created in Table II. We compare the performance of k-map, sample uniqueness and population uniqueness anonymization algorithms. Based purely on anonymization and re-identification risks, K-map would be the best method to use in order to minimize re-identification risk in the loans dataset.

Table III presents the risk assessment for the highest risk identified variables for all three anonymization methods.

TABLE III

High risk features

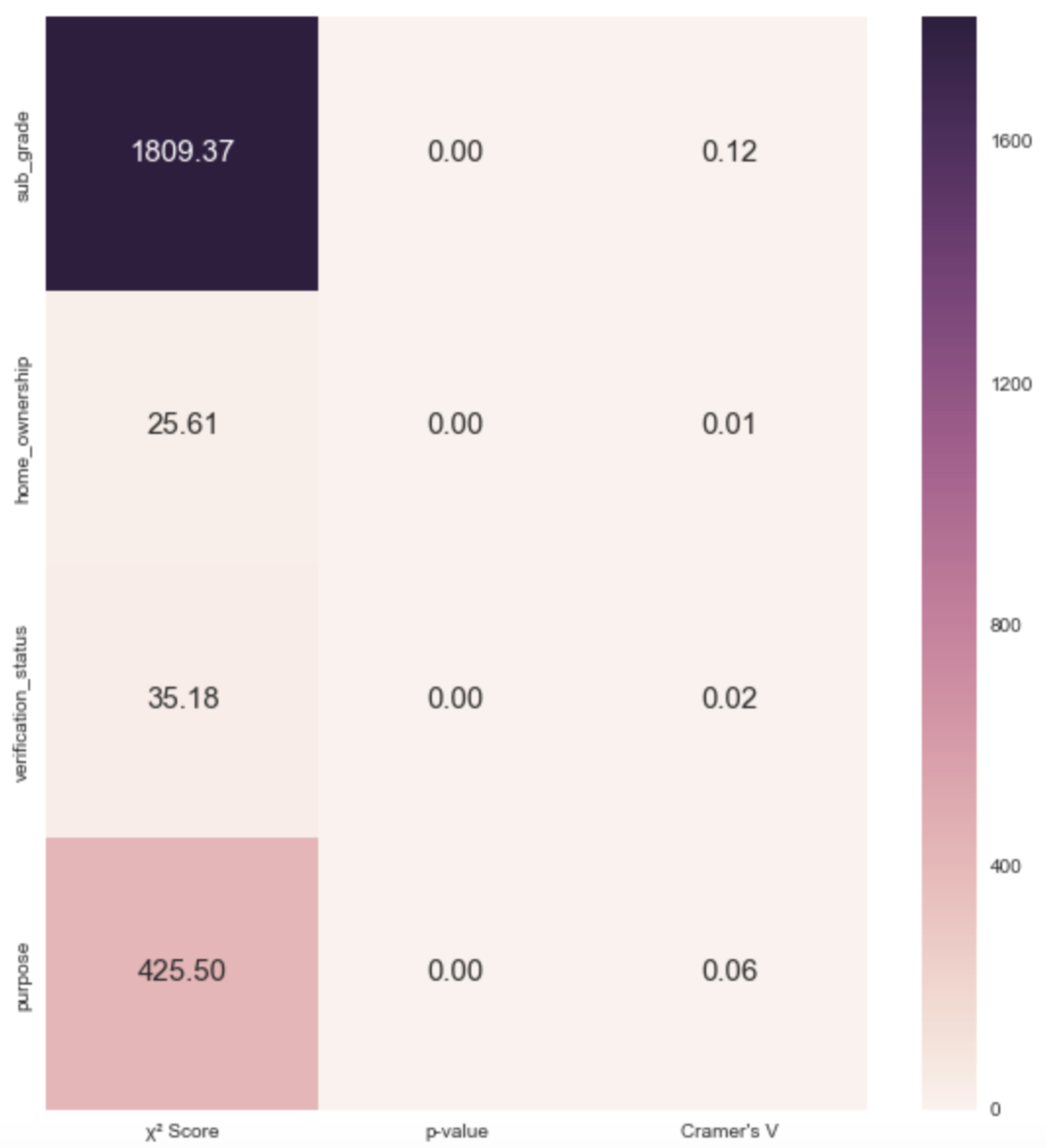


ARX presents an intuitive conclusion: Issue Date, Employment Length and Income are three of the highest risk variables in the loans dataset. All three anonymization algorithms reduce the re-identification risk at the dataset and the feature levels, with K-Map being the most successful anonymization technique.

1. *Data Analysis and Prediction of Loan Outcome*

A chi-squared test of independence on our categorical features shows that all features are significantly related to our dependent variable, loan outcome. The most dependent variable is sub-grade, or the quality of the applicant as established by Lending Club. Effect size, or strength of association, is also strongest (but considered moderately strong) between sub-grade and default at 0.12, which is only a "small" to "medium" effect. Of the categorical features, sub-grade (upper left corner in Fig. 2.2) would be our best predictor of loan outcome:

*Figure 2.2: Chi-squared test of independence categorical features*



Examination of our categorical variables reveals several interesting points to consider for model fitting. Home ownership status is fairly evenly divided between borrowers with a mortgage and borrowers who rent as seen in Fig. 2.3. Income verification seems like it would be very important in the extension of unsecured credit, but Lending Club did not verify income for almost 45% of the loans in this data set. We also determine that debt consolidation is by far the most common borrower purpose for Lending Club loans, seen in Fig. 2.4:

|  |  |
| --- | --- |
| *Figure 2.3: Home Ownership Status* | *Figure 2.4: Loan Purpose for 2007-2011* |
|  | Screen%20Shot%202016-11-15%20at%205.17.25%20PM.png |

Records from the original, pre-processed loan dataset are partitioned into training and test datasets to prepare for logistic regression model fitting. Stratification is employed for splitting data into training and test sets to ensure the distribution of classes in the test set is consistent with the distribution of classes in the entire loans dataset. As an aside, each anonymized dataset is also partitioned into training and test sets consistent with the results of training and test splits for the raw loan dataset. If a record is partitioned into the test set in the after splitting the original loans data, it will also be in the test set for each anonymous dataset. This allows for consistent, accurate comparisons of model results across each dataset. For training and testing purposes, we separate 80% of the records for training and the remaining 20% for testing.

We optimize each logistic regression model using ten-fold cross-validation. Cross-validation is performed on a validation set, which is part of the training set, to estimate generalization performance and tune regularization hyper-parameters. We impose constraints upon the model to control overfitting via regularization hyper-parameters, utilizing both L1 norm (forces most feature weights to zero) and L2 norm (penalizes large individual weights) methods. To determine the most optimal classifier hyper-parameters, we implement grid search cross validation to test different values of the hyper-parameters in a pre-defined space.

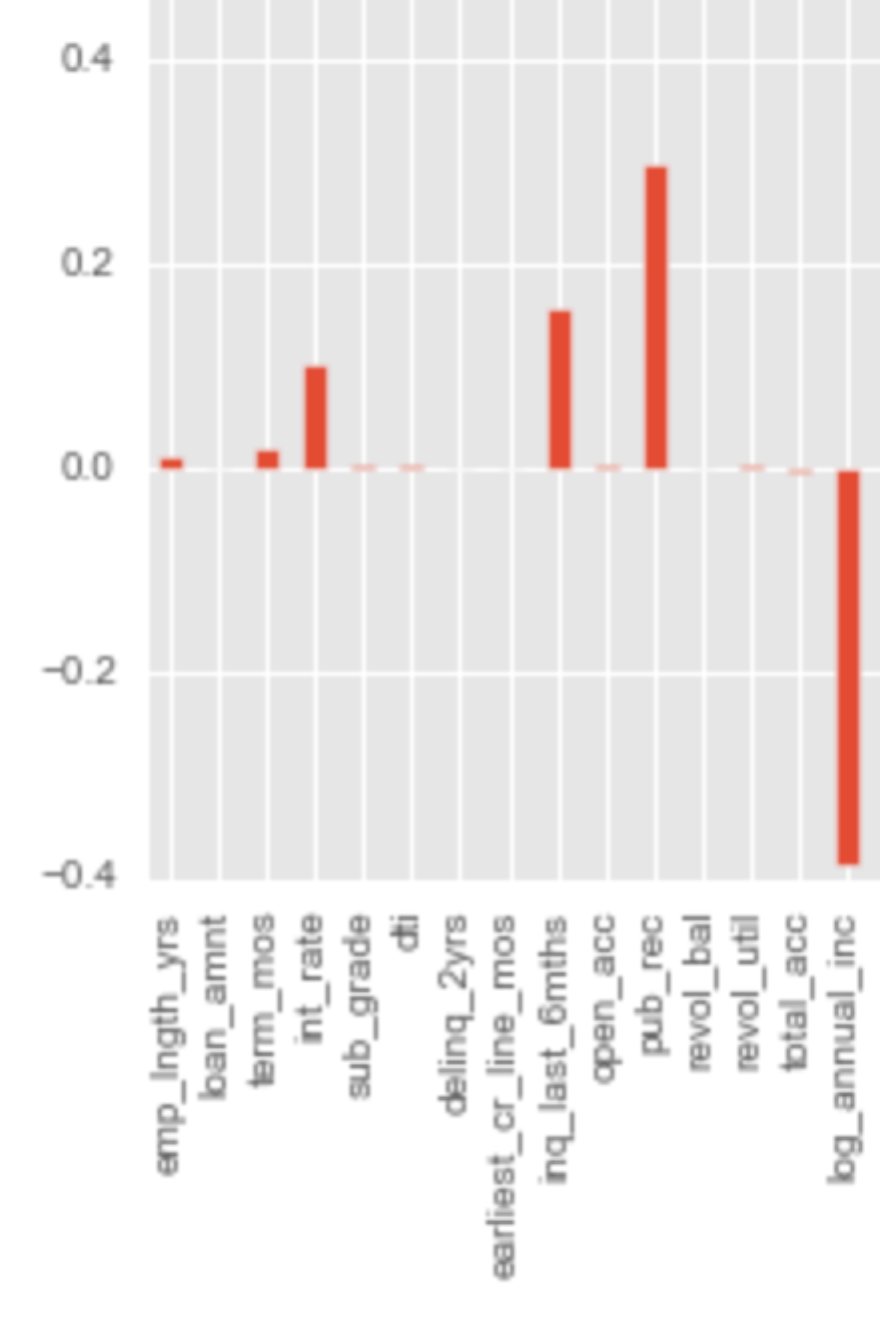
We implement a strategy for managing the imbalance in our binary loan outcome variable (15% “good” and 85% “bad”). Any measures of model accuracy will reflect our underlying class distribution and must be weighted accordingly. We adjust the weight of class 1 (default) relative to class 0 to mitigate the impact of this potential issue.

Classification performance is measured using accuracy and the F1-score. The algorithm that has smaller false positive and negative results is used as best model. F1-score allows us to examine each classification model’s ability to minimize false positive and false negative outcomes.

Maximum F1 and accuracy scores were 65% and 66% for L1 and L2 penalized models on the original loans dataset, respectively. These models can be optimized further for better classification results; however, they provide a good baseline for comparing to anonymized data sets in order to answer our objective question.

We interpret class weights using the coefficients learned via the logistic regression model on the original dataset. For example, since the coefficient of annual income is negative, we can say that an increase in annual income is associated with a decrease in the odds of loan default. This is an intuitive observation for predicting loan outcome:

*Figure 2.5: Logistic Regression Coefficients*



As seen in the logistic regression coefficient breakdown, numerous PII features such as annual income, employment length, and home ownership are present and contain significant coefficient values. These features, especially income, play an important role in our analysis task. However, these features also allow for identification of loan applicants given the right set of companion data, as we have proven previously.

1. *Prediction of Group Membership on the Private Loans Dataset*

The outcome of three different ARX anonymization methods on the ability to run meaningful analysis on our data set can be seen in Table IV. The three methods of anonymization used were population uniqueness, sample uniqueness and k-map anonymization. In order to assess the impact of the anonymization, we perform the same logistic regression method on the anonymized datasets that is performed on the original dataset in order to predict loan outcome.

Similar to the original dataset, the first step is to clean the anonymized data and prepare it for analysis. Each anonymization method obscures different features and each obscured feature is pre-processed for model fitting. Once anonymized features are acceptable for logistic regression, the rest of the dataset is pre-processed in a similar manner to the original dataset, using the same methods for imputation and deletion.

The dataset that was anonymized using the population uniqueness method obscures the loan\_amnt, issue\_d, earliest\_cr\_line\_mos, and revol\_bal variables. The issue\_d variable was recoded as a decade. Given that most of the loans in the data were issued in the same decade, we decided to drop the issue\_d variable for this dataset. The remaining anonymized variables were converted to a leading integer with the remainder of the field replaced by asterisks. We strip off the trailing asterisks and included the leading integers in our model building process.

With the sample uniqueness anonymization method, the loan\_amnt, earliest\_cr\_line\_mos, and revol\_bal fields are obscured. Each of these fields is replaced completely with asterisks. Given that none of these fields contains usable data any longer, we decide to drop them all from the dataset.

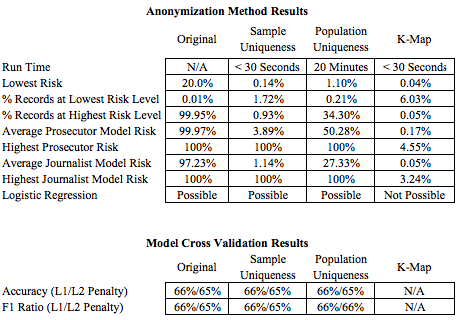
Using the K-Map anonymization method, the loan\_amnt, home\_ownership, loan\_status, earliest\_cr\_line\_mos, and revol\_bal columns are all obscured. The response, loan outcome, in this dataset is built off of the loan\_status variable. Given that the loan\_status variable was completely obscured by the k-map anonymization process, we are unable to perform logistic regression on this version of the data to predict whether or not a loan will default.

Performing logistic regression on the dataset anonymized using the population uniqueness method, we see similar results compared to the original dataset. Both L1 and L2 penalties result in very similar results. When using balanced weights, the results are nearly identical to the results of the same model applied to the original dataset. Anonymization using population uniqueness does not impact the ability to classify loan outcome. The model with L1 norm penalty fit on the population uniqueness dataset produces average accuracy and F1 scores of 0.65 and 0.66, respectively, across all 10 folds of cross validation. The L2 regularization model produces an average score of 0.66 for both accuracy and F1.

When using logistic regression with balanced weights on the dataset anonymized using sample uniqueness, results are similar to the to the original and population uniqueness anonymized dataset results. The L1 regularization model fit on this dataset produces an average score of 0.66 for both accuracy and F1, while the L2 regularized model produces an average score of 0.65 for both accuracy and F1.

TABLE IV

Re-identification risk: Analysis Results



1. *Model Comparisons Anonymization’s Impact on Data Analysis*

The results from logistic regression models fit on the original loans dataset and two of the three anonymized datasets all appear to be similar. In order to investigate this result statistically, we run a one-way ANOVA on the accuracy and F1 for the models using both L1 and L2 penalties. When validating the logistic regression models, we store the accuracy and specificity results for each fold of cross validation in an array. These arrays act as the inputs for the one-way ANOVAs that were run for each penalty level.

The null hypothesis for each ANOVA is that the means of each given statistic for the three populations are equal. We use an alpha of 0.05:

TABLE V

p-Values of One-Way ANOVA Tests

|  |  |  |
| --- | --- | --- |
|  | Accuracy | F1 |
| L1 Regularization | 0.215 | 0.186 |
| L2 Regularization | 0.810 | 0.565 |

As Table V illustrates, the p-values for accuracy and F1 for each model are insignificant. Therefore, we cannot reject the null hypothesis that the population means are equal for any of them. There is no statistically significant evidence that the population uniqueness and sample uniqueness anonymization techniques used on the loans dataset impact the ability to create statistical models to predict loan outcome. The K-map anonymization technique is the only method used that impacted data analysis.

V. FINAL OBSERVATIONS

While the k-map anonymization technique did not provide for an appropriate dataset to analyze, both the sample uniqueness and population uniqueness methods resulted in classification results similar to using the original, de-anonymized dataset.

For the specific task of classifying loan outcome for the Lending Club loans dataset, we illustrate that data anonymization does not affect analysis outcomes in two out of three results. However, given the variability of analysis tasks and datasets, results can vary drastically. Our results cannot be extrapolated to the entire population of data analysis tasks or datasets used for analysis.

Another key limitation of our analysis is that both the population uniqueness and sample uniqueness anonymization methods resulted in reduced, but still apparent re-identification risk. Anonymization should be attempted given data that provides for personal re-identification or contains PII features.

Public datasets be anonymized according to regulatory standards such as the HIPAA Privacy standards. Anonymization techniques are not always successful and data removal should be considered if privacy risks are a concern.

REFERENCES

1. Schneier, B. “Anonymity and the Netflix Dataset.” (2007, Dec.) [Online] Available: <https://www.schneier.com/blog/archives/2007/12/anonymity_and_t_2.html>
2. Lande, W., Rao, R. (2003, Jan.). “Secure De-identification and Re-identification.” Computer Aided Diagnosis & Therapy Group, Siemens Medical Solutions. Malvern, PA., United States. [Online]. Available:

<https://www.media.mit.edu/research/highlights/unique-shopping-mall-reidentifiability-credit-card-metadata>

1. Montjoye, Y. (2015, Jan.). “Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata” MIT Media Lab. Cambridge, MA., United States. [Online]. Available:

<https://www.media.mit.edu/research/highlights/unique-shopping-mall-reidentifiability-credit-card-metadata>

1. “Summary of the HIPAA Privacy Rule.” U.S. Department of Health and Human Services. United States. [Online]. Available:

<https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/>

1. No Author. (2014, Mar.). “Overview of Data Anonymization.” [Online]. Available:

<http://www.odbms.org/wp-content/uploads/2014/03/The-Complete-Book-of-Data-Anonymization_Chap_1.pdf>

1. Office for National Statistics. (2013, Jul.). “Beyond 2011: Matching Anonymous Data.” United Kingdom. [Online]. Available: <http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-matching-anonymous-data--m9-.pdf>
2. Narayanan, A., Shmatikov, V. (2007, Nov.). “Robust De-Anonymization of Large Sparse Datasets.” University of Texas at Austin. Austin, TX., United States. [Online]. Available:

<https://www.cs.cornell.edu/~shmat/shmat_oak08netflix.pdf>

1. ARX De-identifier. [Online]. Available: <http://arx.deidentifier.org/>
2. Nelson, G. “Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification.” (2015). Thotwave Technologies. Chapel Hill, NC. [Online]. Available: <http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf>
3. Facebook Graph API. Facebook Inc. [Online]. Available: <https://developers.facebook.com/docs/graph-api>
4. Sweeney, L. “Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4.” Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000
5. Federal Governmenet Zip Codes. [Online].

Available: <http://federalgovernmentzipcodes.us/>

1. Ghose, T. (2013, Apr.) “Think You’re Anonymous? Cell Phone Data Gives You Away.” [Online]. Available: <http://www.livescience.com/28353-anonymous-phone-data-not-anonymous.html>
2. D. Hsu. (2015, Apr.) “Techniques to Anonymize Human Data.” Datasift. United States. [Online]. Available:<http://blog.datasift.com/2015/04/09/techniques-to-anonymize-human-data/>
3. E. McAllister, T. Grance, K. Scarfone. (2010, Apr.). “Guide to Protecting the Confidentiality of Personally Identifiable Information (PII).” National Institute of Standards and Technology. Gaithersburg, MD., United States. [Online]. Available: <http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf>
4. W. Kan. (2016, Jun.). Lending Club Loan Data, 2016. Lending Club. United States. [Online]. Available: <https://www.kaggle.com/wendykan/lending-club-loan-data>
5. “Rules and Policies – Protecting PII – Privacy Act.” U.S. General Services Administration. United States. [Online].

<http://www.gsa.gov/portal/content/104256>

1. Emam, K. Dankar, F. (2008, Oct.) “Protecting Privacy using K-Anonymity.” [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/>

1. A Data Tug of War: Privacy and Data Analysis is submitted for review by Dr. Daniel Engels on Dec. 9, 2016. This work was supported in part by the MSDS program at SMU.

   Anise Mulkey is a student in the SMU MS Data Science Program from Dallas, Texas (e-mail: amulkey@mail.smu.edu)

     Zach Brown  is a student in the SMU MS Data Science Program from Boston, Massachusetts (e-mail: [zbrown@mail.smu.edu](mailto:zbrown@mail.smu.edu))

     Kumar Raja Guvindan Raju is a student in the SMU MS Data Science Program from San Diego, California (e-mail: [kguvindanraj@mail.smu.edu](mailto:kguvindanraj@mail.smu.edu))

     Cory Nichols is a student in the SMU MS Data Science Program from Austin, Texas (e-mail: [pcnichols@smu.edu](mailto:pcnichols@smu.edu)) [↑](#footnote-ref-1)