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

[[1]](#footnote-1)

Anonymous Prediction: An Empirical Comparison

***Abstract* – To be completed**.

T

I. INTRODUCTION

HE value of personally identifiable information is undeniable. Marketers can tailor specific 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 today. Digging even deeper, These features can be married to other public datasets or used with statistical analysis methods to personally identify individuals with high confidence. Given an extensive set of features for a patient, medical information can easily be triangulated to identify a person, and subsequently, the person’s medical history. This information can lead to additional marketing or even worse: unwanted health status exposure to public entities, employers or hackers. Too often, datasets used for research contain features about individuals or entities that allow for the compromise of individual anonymity.

A broad solution is to ensure full data anonymity in datasets used for research and analysis.

Brute force methods such as data truncation or outright data exclusion have been utilized in the past to achieve the anonymity of personally identifiable information in datasets containing people as objects or rows. Telephone numbers, addresses, social security numbers and credit card information are outright deleted from the dataset or truncated to a number of digits that represent an abstraction of the original data. However, the effects of brute force anonymization techniques, especially the deletion of features, can severely impact the results of data analysis.

Striking a balance between maintaining anonymity and obtaining meaningful analytical results is a formidable challenge. Our analysis suggests there is a “sweet spot” between ensuring anonymity and obtaining meaningful analytical results. More specifically, we determine how much a dataset needs to be modified in order to be considered anonymized, while still allowing for meaningful analytical results.

For our analysis, we utilize a public dataset of personal loans retrieved from Kaggle.com. Our dataset contains 887,379 unique records or objects, with a mixture of 74 categorical and numeric features including features describing the loan’s interest rate, home ownership status, employment title and annual income.

THE DELETED STUFF IS TOO MUCH DETAIL, ESPECIALLY FOR AN INTRODUCTION. WHAT YOU DID NEEDS TO BE A HIGH LEVEL DESCRIPTION. GIVE THE OVERVIEW NOT THE DETAILS. AND YOU NEED TO GIVE THE RESULTS.

Main results will go here.

Our paper is organized as follows. IN SECTION II WE… IN SECTION III WE….

REFERENCES

1. Kan, Wendy. (June, 2016). Lending Club Loan Data, 2016. Lending Club. United States. [Online]. Available: <https://www.kaggle.com/wendykan/lending-club-loan-data>
2. McAllister, Erika, Grance, Tim, Scarfone, Karen. (April, 2010). 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>
3. Hsu, Dary. (April, 2015) Techniques to Anonymize Human Data. Datasift. United States. [Online]. Available:<http://blog.datasift.com/2015/04/09/techniques-to-anonymize-human-data/>
4. Nelson, Gregory. (2015). Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. Thotwave Technologies. Chapel Hill, NC. [Online]. Available: <http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf>
5. Overview. ARX Data Anonymization. Retrieved September 12, 2016 from

<http://arx.deidentifier.org/overview/>

1. Anonymous Prediction: An Empirical Comparison is submitted for review by Dr. Daniel Engels on Sep. 6, 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)