Poster: Privacy and De-Identification in High-Dimensional Social Science Datasets

A. Cheyenne Solomon*, Raquel Hill*, Erick Janssen**, and Stephanie Sanders**
** School of Informatics and Computing, Indiana University*
*** Kinsey Institute for Research in Sex, Gender, and Reproduction, Indiana University*

***Abstract*—Social scientists often gather large amounts of sensitive data. Unlike some medical-related datasets, these social science datasets tend to be sparse and high-dimensional. This dimensionality increases the possibility that participants in the dataset provide answer patterns that characterize them in unique ways. Although at this stage the vulnerabilities involved remain to be established, it is possible that these unique characterizations enable individuals to be linked to external data in ways that may not have been previously considered. Thus, 'uniquifiability' may increase identifiability, and this may mean that traditional approaches to de-identifying data, such as fulfilling HIPPA requirements, may not be sufficient for preventing the re-identification of participants in large social science datasets.**

In this project, we evaluate the statistical characteristics of a large social science dataset to better understand how unique features impact privacy. Our preliminary results show that 36% of the participants within the dataset are unique even when considering only one data attribute.

I. Introduction

Datasets such as hospital records tend to include very small numbers of fields, such as name, date, complaint, and diagnosis. Social science datasets, including those collected by the Kinsey Institute for Research in Sex, Gender, and Reproduction, have comparatively high-dimensional data. Each participant may be asked to answer many questions on various surveys, increasing the amount of information available per individual, and making it easier to correlate attributes to extract data and identify individuals. Our work and main contribution focuses on determining what Kinsey data fields an adversary could use to re-identify survey participants. We do this in order to understand the properties of the data so that we can find the best way to protect it and the participants who provided it and maintain the utility for which the data were gathered. Although our analyses focus on one specific type of dataset, relevant to sexual health research, we hope that our current approach is applicable to social science datasets in general, and we will be continuing work in that direction.

The conventional wisdom for anonymizing data to protect an individual's privacy is that if we remove certain pieces of information, data cannot be linked to a specific person. The current standard for preserving the privacy of participants within human health data is HIPAA Safe Harbor. To avoid stringent authorization requirements of data sharing and usage, Safe Harbor requires that data be 'de-identified' by removing 18 identifiers [4]. These identifiers include birthdates and any age information for those over 89, geographic locations finer than state, various identifying numbers such as license or insurance plan numbers, emails, IPs, facial photographs, etc. After data have been de-identified per Safe Harbor, they can be processed and shared based on the assumption that doing so does not expose subjects to re-identification attacks.

In addition to the 18 identifiers, HIPPA Safe Harbor also requires the removal of "any other unique, identifying characteristic or code". We propose that almost any type of information could fall under that last provision. Outside knowledge has been found to be a threat to 'anonymized' datasets [2]. Prior anonymity research, such as k-anonymity [3], assumes that an adversary trying to identify a participant within a dataset has access to very limited outside knowledge, such as voter registration records or other publicly-available government datasets. l-diversity [1] more directly addresses questions relevant to outside knowledge, but still limits it to very identifying information in the same vein as public datasets. However, in the age of social networks and rampant self-reporting of sensitive and/or identifying information, such assumptions are unfounded.

II. Results

Preliminary results using a Kinsey Institute dataset suggests that almost any attribute can be a quasi-identifier; that is, in combination with other attributes, a participant can be found to be unique. Even answers to simple and general multiple choice questions can be combined to uniquify[1] participants.

We analyzed the data to answer the following questions:

- How many participants are unique in various combinations of fields?

[1] The term 'uniquifiy' is introduced here in reference to the reduction of the query population to a single participant. A participant's answers to a survey question or set of survey questions is unique if he or she is the only person with those specific answers.

- How many fields must be combined before 100%, or nearly 100%, of participants have at least 1 unique answer or combination of answers?
- What fields make participants more vulnerable than others?
- Are text-entry fields more likely to lead to unique response patterns than multiple-choice fields?
- Does partitioning the fields, such as by survey module, affect participant uniquifiability?
- Which participants are unique many times and what is it about their data that makes them that way?
- Can uniquification predict other factors about a participant?[2]

Table 1 summarizes our preliminary results and shows that even when considering only two data fields, over 50% of the participants within the dataset are unique. Many times participants have more than twenty separate unique answer combinations. To determine uniqueness, we compare each participants' answers to every other participants' answers. Combinations are unique if the combination of both answers is unique, not if both in the combination are singly unique. Answers for which a participant is singly unique are not considered in combination analysis because all combinations with such answers would be trivially unique.

III. Continuing and Future Work

Our analyses are ongoing and we have analyzed an additional Kinsey dataset and the effects of dataset partitioning. We have looked into and will continue to investigate how survey development, such as trying to get normally distributed results, affects uniquification rates. We will also look at combinations of three answers with an eye towards a possible 100% uniquification rate.

In the future, we will explore whether unique features may increase the probability of re-identification of individuals. We believe that, at the very least, people accurately self-report their identities online, making access to more reliable government databases unnecessary. If one can connect data from an accessible Kinsey dataset to an online identity, such self-reported data could provide an avenue for re-identification of survey participants. Additionally, we plan to investigate whether our results will be applicable to other social science datasets and fields. If social science datasets share re-identification properties, we can develop protection models and mechanisms that can be applied broadly.

IV. Acknowledgements

This work is funded by NSF grant CNS-101 2081.

References

[1] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1, March 2007.

[2] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. *IEE Symposium on Security and Privacy*, 0:111–125, 2008.

[3] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10:557–570, October 2002.

[4] U.S. Department of Health and Human Services. HIPAA Administrative Simplification, 2006. Section 164.514 (b): Implementation specifications: Requirements for de-identification of protected health information.

Analysis	Percentage
Uniquified by 1 answer	36.31
Uniquified by 1 answer excluding text entry answers*	3.76
Uniquified by 2-answer combinations	79.27
Uniquified by 2-answer combinations excluding text-entry answers*	54.61
Uniquified by 2-answer combinations, but not 1 answer	42.96
Uniquified in more than 100 ways	14.88
Uniquified in more than 20 ways	36.52

* Text-entry answers are most often trivially unique, e.g. "Aderall, Advaire" could be unique due to misspellings. Omitting them, we believe, gives a more accurate picture of uniquification.

Table I: Initial results of simple data analyses. There are a total of 10866 participants and 332 fields, each of which is an answer to a survey question. Participants' answers were compared to all others' to determine uniquiness. Two-answer combinations are unique if no other participant has that combination, but the participant need not be unique in both singly. Note that singly unique answers were filtered out for combination analysis, as such answers would make participants trivially unique in all combinations with singly unique answers.

[2] This question is more on the social science side, as it refers to behavioral predictions, but has implications for privacy as well.