Is Your Privacy Lost in *Transition*? Analyzing Transitive Disclosure Risks in Open Datasets

Kaustav Bhattacharjee[1], Akm Islam[1], Aritra Dasgupta[2]
[1] Department of Informatics, New Jersey Institute of Technology
[2] Department of Data Science, New Jersey Institute of Technology
{kb526, azi3, aritra.dasgupta}@njit.edu

***Abstract*—Government agencies collect citizens' data, process them and release them online after de-identification so that this data can be used for research purposes. Recent studies have shown that a pair of these open datasets can be joined together to carry out linking attacks, hence data custodians lookout for the joinability risk between two datasets. However, an adversary can exploit the transitive closure property and use multiple datasets to join two seemingly un-joinable datasets. This can lead to the disclosure of sensitive information about the consumers whose data have been captured in these datasets. Hence, we propose to develop a visual analytic framework to assess the transitive disclosure risk of open datasets and evaluate them for actual disclosures through a web-based interface.**

***Index Terms*—open datasets, disclosure risk, data subjects, data privacy, data custodians**

I. Motivation

Large-scale data collection has now been adopted by many government agencies. These agencies collect citizens' data, de-identify them and release them online as open datasets for public consumption. Portals like the NYC Open Data portal [1] and the City of Dallas Open Data portal [2] serve as websites where a broad spectrum of users like researchers, analysts, and even adversaries can access datasets of interest. The open data portals are generally inspired by the FAIR data principles [3], which promote the findability, accessibility, interoperability, and reuse of proprietary information. But on the other hand, these portals may be prone to inadvertent data leaks, causing irreplaceable damage to the data subjects mentioned in these portals. A recent study has shown that researchers were able to re-identify the details for 91% of all the taxis in NYC using an anonymized open taxi dataset and an external medallion dataset [4]. In another example, researchers re-identified individuals from de-identified medical records for the 10% of the population, released by the Australian Department of Health [5]. Lavrenovs and Podins showed that the passengers' private information could be disclosed through the public transportation open data released by the city municipal of Riga, Latvia [6]. Another study showed that 99% of Americans can be re-identified even from heavily anonymized and incomplete datasets using a combination of a few demographic attributes [7]. Inspired by these examples and the concept of transitive dependency in databases [8], we would like to explore the concept that two datasets, which have no shared attributes between them, can still be joined if they have shared attributes with a third dataset. In this proposal, we focus on assessing the risk of disclosure through transition (or *transitive disclosure risk*) in open datasets, in order to prevent the disclosure of sensitive information about an individual or a group of individuals.

II. Approach

To analyze the transitive disclosure risk, we propose the development of a visual analytic framework along with a web-based interface targeted toward data custodians. Some of the possible steps of our approach are as follows:

- *Curation of privacy-relevant datasets:* Open data portals can be considered as vast repositories of datasets that may or may not be relevant to users in the context of privacy. We plan to create a set of highly relevant datasets and a rich taxonomy of their metadata like domain tags, granularity, attributes, data portals, and others.
- *Development of a graph of dataset joinability:* Datasets that can be joined based on shared attributes might disclose sensitive information about data subject(s). However, in the forest of open datasets, it is difficult to analyze each and every combination of the datasets for joinability. Hence, we plan to create a graph of these datasets where the edges represent the joinability between them. This will help us discover potential transitive paths between the datasets which are otherwise not joinable directly.
- *Assessment of transitive disclosure risk:* Using the graph of dataset joinability and the types of joins possible, we plan to assess the risk associated with each dataset and subsequently develop a metric for it. This transitive disclosure risk metric will help the users triage the datasets effectively.
- *Development of a visual analytic framework and interface:* Assessing transitive disclosure risk and evaluating the datasets for actual disclosures may be an iterative process. Though some of the steps can be automated, there may be multiple transitive dataset combinations. Each of them may lead to different degrees of disclosures, which need to be evaluated by a human expert. Visual analytic interventions can help improve this human-in-the-loop endeavor; hence, we plan to develop a framework that researchers can use to streamline the entire process. This framework will be used to implement an interface that will help users effectively analyze the transitive disclosure risks and evaluate them for disclosures.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] "NYC Open Data." https://opendata.cityofnewyork.us/. (Accessed on 10/05/2021).

[2] "City of Dallas Open Data." https://www.dallasopendata.com/. (Accessed on 10/05/2021).

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[4] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing nyc taxi data: Does it matter?," in *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pp. 140–148, IEEE, 2016.

[5] C. Culnane, B. I. Rubinstein, and V. Teague, "Health data in an open world," *arXiv preprint arXiv:1712.05627*, 2017.

[6] A. Lavrenovs and K. Podins, "Privacy violations in riga open data public transport system," in *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, pp. 1–6, IEEE, 2016.

[7] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[8] E. F. Codd, "Further normalization of the data base relational model," *Data base systems*, vol. 6, pp. 33–64, 1972.