3-13-2019

Preserving privacy in data analytics

Lila Ghemri
Texas Southern University

Preserving Privacy in Data Analytics

Lila Ghemri,
Texas Southern University
Houston, Texas, USA
lila.ghemri@tsu.edu

ABSTRACT

Data Analytics is becoming an essential business tool for many data intensive companies and organizations. However, the increased use of such methods comes with the threat of data disclosure. Privacy-preserving methods have been developed with varying degrees of efficiency with the main goal of protecting individuals' privacy. This tutorial aims at presenting models and techniques of preserving privacy in machine learning and data mining.

KEYWORDS

Privacy-preserving methods, privacy, data analytics, tutorial.

ACM Reference format:

Lila Ghemri. 2019. Privacy Preserving in Data Analytics. In *Fifth ACM International Workshop on Security and Privacy Analytics (IWSPA '19). March 27, 2019, Richardson, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3309182.3311786

1 INTRODUCTION

Concerns about data protection and privacy have been at the forefront of distributed data base designers for decades [3, 11]. These concerns have only exacerbated with the advent of online social networks, e-commerce, smart devices and software as a service using the cloud. Rarely does a month go by without the media reporting a data breach in which thousands or millions of customers' personal data, such as Social Security Numbers, credit card numbers, etc., have been disclosed [7]. The sheer amount of compromised data is a result of companies and organizations, such as Amazon and Google, collecting and storing data related to their business and customers/users in order to gain insight on how to improve their users' experience. These data are usually uploaded to the cloud and used by data analytics programs to train learners and gain business insights and marketing advantage. While users may benefit from such efforts by enjoying a better user experience, collecting and processing users' data can constitute a threat to their privacy in that it may potentially reveal sensitive information about them, such as their spending habits, gender preferences, or any other information they deem private. Several methods have been developed to enable the collection and processing of potentially sensitive data in a privacy-preserving manner. Most of these methods rely on procedures for data perturbation that mask either the original data or the processing results, to protect from disclosure.

Privacy-preserving methods can be applied pre data analytics in which case, the dataset is modified before it is released to the data analytics component. These methods rely on various data perturbation techniques which satisfy some statistical constraints. Alternatively, they can be applied post data analytics, in which case the returned results are "anonymized" using output perturbation techniques, (refer to Aggarwal and Yu for extensive coverage [1]).

2 PRIVACY VERSUS UTILITY

A natural consequence of applying Privacy-Preservation is the loss of information. The loss of information on specific data items may not only affect the quality of the dataset but also that of the results obtained from the data analytics component. Hence, utility-based privacy preservation techniques have been developed with two goals in mind: protecting the private information and as much as possible preserving data utility [9].

The aim of this tutorial is to present an overview of privacy threats, methods used for privacy-preserving data analytics and examples of their applications across multiple domains.

3 TUTORIAL OUTLINE

This tutorial will introduce the audience to the following aspects of privacy-preserving data analytics:

1) The notion of privacy
2) Attributes and Identifiers
3) Types of privacy attacks:
    - "Good credit/Bad credit" a.k.a. Leakage attacks
    - "Who gave you my name?" a.k.a. Linkage attacks
    - "I know where you have been" a.k.a. Trail Re-identification attacks

IWSPA '19, March 27, 2019, *Richardson, TX, USA*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6178-1/19/03…$15.00
https://doi.org/10.1145/3309182.3311786

- "So you got a pay raise!" a.k.a. Tracker attacks

4) Randomization Techniques
5) Grouping and Data Partitioning Techniques
6) Data Obfuscation and Differential Privacy
7) Inference Control
8) Secure Multi-party Computation
9) Attacks techniques on Privacy-Preserving Methods
10) Privacy approaches for specific machine learning techniques such as Neural networks, regression analysis, etc.[2, 4, 5, 8]

4 TEACHING PRIVACY

An extended version of this tutorial has been developed and has been presented as part of a senior and a graduate database course and a data mining/machine learning course. The extended version consists of three parts: The Concept of Privacy, Data Perturbation Methods, Mining Association Rules and Distributed Data under Privacy Constraints. The parts are independent and have been presented several times. Student surveys show that these modules have provided them with an increased awareness of the importance of the topic of privacy and the techniques available to protect user's data [6, 10]

5 SHORT BIOGRAPHY OF TUTORIAL PRESENTER

Lila Ghemri is a Professor of Computer Science at Texas Southern University located in Houston, Texas. She received her Ph.D. in Computer Science from Bristol University in the UK. Ghemri spent ten years in research and industry developing NLP and knowledge based systems before joining academia. Her research interest include security and privacy in medical data, online social networks and mobile applications.

ACKNOWLEDGMENTS

This work was partially supported by NSF under grants DUE 1241626 and the DHS Command, Control and Interoperability Center for Advanced Data Analysis. The author would like to thank Rakesh Verma for his comments and suggestions.

REFERENCES

[1] Charu C. Aggarwal and Philip S. Yu (Ed). 2008. Privacy-Preserving Data Mining: Models and Algorithms. Springer, USA. DOI:10.1007/978-0-387-70992-5.

[2] Ahmed AlEroud. 2017. Anonymization of Network Trace Using Differential Privacy. RIPE 74. Budapest, Hungary, May 8-12, 2017.

[3] Jan Camenish and Anja Lehmann. 2015. (Un)linkable Pseudonyms for Governmental Databases. In Proceedings of the 22nd, ACM SIGSAC Conference on Computer and Communications Security (CCS '15). ACM, New York, NY, USA, 1467-1479.

[4] Giulia Fanti, Vasyl Pihur, and Úlfar Erlingsson. 2016. Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries. *In Proceedings on Privacy Enhancing Technologies ;* 2016 (3):1–21

[5] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. 2017. Privacy-Preserving Distributed Linear Regression on High-Dimensional Data. *In **Proceedings** on Privacy Enhancing Technologies.* 2017 (4):345–364

[6] Lila Ghemri and Ping Chen. 2013. Introducing privacy in a data mining course (abstract only*). In* Proceeding of the 44th ACM technical symposium on Computer science education *(SIGCSE '13).* ACM, New York, NY, USA, 740-740. DOI: https://doi.org/10.1145/2445196.2445445

[7] Leonid Grustniy. 2017. Top 5 Largest Data Leaks in 2017- so far. Kaspersky Labs. https://www.kaspersky.com/blog/data-leaks-2017/19723/

[8] Ehsan Hesamifard, Hassan Takabi, Mehdi. Ghasemi and Rebecca .N. Wright**.** Privacy-preserving Machine Learning as a Service**.** *In Proceedings on Privacy Enhancing Technologies***,** 2018**(**3**),** 123–142**.**

[9] Ming Hua. 2008. A Survey of Utility-Based Privacy-Preserving Data Transformation Methods, Privacy- Preserving Data Mining: Models and Algorithms. Springer, USA. DOI: 10.1007/978-0-387-70992-5.

[10] Ernst Leiss and Lila Ghemri. 2014. Privacy between technological capabilities and society's expectations (abstract only*). In* Proceedings of the 45th ACM technical symposium on Computer science education *(SIGCSE '14).* ACM, New York, NY, USA, 727-727.

[11] Michael Siegenthaler and Ken Birman. 2009. Privacy enforcement for distributed healthcare queries. *In Proceedings of the3rd International Conference on Pervasive Computing Technologies for Healthcare. (PervasiveHealth 2009).* 1-6, DOI: 10.1109/NCA.2009.33