Privacy-preserving Data Mining in Industry

Krishnaram Kenthapadi
LinkedIn
kkenthapadi@linkedin.com

Ilya Mironov
Google
mironov@google.com

Abhradeep Guha Thakurta
UC Santa Cruz
aguhatha@ucsc.edu

ABSTRACT

Preserving privacy of users is a key requirement of web-scale data mining applications and systems such as web search, recommender systems, crowdsourced platforms, and analytics applications, and has witnessed a renewed focus in light of recent data breaches and new regulations such as GDPR. In this tutorial, we will first present an overview of privacy breaches over the last two decades and the lessons learned, key regulations and laws, and evolution of privacy techniques leading to differential privacy definition / techniques. Then, we will focus on the application of privacy-preserving data mining techniques in practice, by presenting case studies such as Apple's differential privacy deployment for iOS / macOS, Google's RAPPOR, LinkedIn Salary, and Microsoft's differential privacy deployment for collecting Windows telemetry. We will conclude with open problems and challenges for the data mining / machine learning community, based on our experiences in industry.

ACM Reference Format:
Krishnaram Kenthapadi, Ilya Mironov, and Abhradeep Guha Thakurta. 2019. Privacy-preserving Data Mining in Industry. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3289600.3291384

1 OUTLINE OF THE TUTORIAL

1.1 Privacy Breaches and Lessons Learned

We will present the key privacy breaches over the last two decades, highlighting the attacker's advantage and the lessons learned in each case. The examples to be presented include the following.

- Sweeney's de-anonymization of MA governor's health records [25] (Lesson learned: The attacker can make use of auxiliary information (which can sometimes come from unexpected sources).)
- AOL search log release [2] (Lesson learned: Significant damage can be inflicted even if the attacker is able to succeed on a small fraction of the inputs.)
- De-anonymizing Netflix data [20] and de-anonymizing web browsing data with social networks [24] (Lesson learned: These attacks highlight the high dimensionality of the data, which also makes them robust/scalable.)
- Privacy attacks on microtargeted ads [18] (Lesson learned: The attacker can play an active role, and can choose to be part of the dataset (e.g., via creating fake user profiles), and/or have an influence on the creation of the dataset (e.g., by setting up appropriate microtargeted ads).)
- Privacy attacks on collaborative filtering [6] (Lesson learned: The attacker can observe how the system changes over time (or other dimensions).)

1.2 Key Privacy Regulations and Laws

We will give a brief overview of key privacy regulations and laws such as GDPR (General Data Protection Regulation) [26], which took effect in May, 2018 as well as earlier laws such as HIPAA (Health Insurance Portability and Accountability Act).

1.3 Differential Privacy: Definition and Properties

We will motivate the need for a rigorous privacy guarantee, and present the notion of differential privacy [8–10]. Differential privacy is a formal guarantee for preserving the privacy of any individual when releasing aggregate statistical information about a set of people. In a nutshell, the differential privacy definition requires that the probability distribution of the released results be nearly the same irrespective of whether an individual's data is included as part of the dataset. As a result, upon seeing a published statistic, an attacker would gain very little additional knowledge about any specific individual. We will also give different interpretations of differential privacy, and its properties.

1.4 Privacy Techniques in Practice: Challenges and Lessons Learned

The privacy attacks over the last two decades have highlighted the need for adopting rigorous privacy techniques and demonstrated that it is highly non-trivial to balance the trade-offs between utility and privacy. Utility vs. privacy trade-offs have been studied in the literature, but handling the trade-offs between computation and communication resources is unique to industrial deployment. Current industrial deployments are highly distributed. Each device (e.g., a cell phone) holds a single data point, and the server computes aggregates over differentially private information obtained from the devices. Although server computation is cheap, client computation is expensive since the clients are usually low-power devices. Furthermore, client / server communication is expensive, and hence both the amount of communication and the number of rounds of communication have to be minimized.

Lessons learned: The notion of differential privacy is a principled foundation for privacy-preserving data analyses, which has been witnessing practical adoption in industry as well as in government (e.g., by U.S. Census Bureau [14, 19]). In particular, local differential privacy is a powerful technique suitable for internet-scale telemetry. While applying these techniques in practice, we have learned that resource constraints drive the algorithmic design as much as the

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02.
https://doi.org/10.1145/3289600.3291384

privacy vs. utility trade-offs. In fact, often these constraints enable the design of theoretically optimal algorithms (e.g., [3]).

1.5 Case Studies in Industry

As part of the tutorial, we will also focus on the application of privacy-preserving data mining techniques in practice, by presenting case studies from different technology companies (including Google, Apple, LinkedIn, and Microsoft).

1.5.1 Case Study: Google's RAPPOR. We will describe Google's RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response) project [11, 12], which has been the first large scale deployment of differential privacy in industry. RAPPOR is built on the concept of randomized response, and enables a privacy-preserving way to learn software statistics (e.g., about how unwanted software is hijacking users' settings in Chrome browser) that can be used to improve browser security, find bugs, and provide better overall user experience. RAPPOR has been made available as an open-source project (https://github.com/google/rappor).

1.5.2 Case Study: Apple's Local Differential Privacy Deployment for iOS / macOS. We will present an overview of Apple's local differential privacy deployment for iOS and macOS, highlighting the application scenarios and the design choices [1].

Algorithm Design: We will describe the detailed algorithm based on the following patents (https://www.google.com/patents/US9594741 and https://www.google.com/patents/US9705908), and the follow-up paper [3]. The high-level idea is to use techniques from sketching and data streaming literature and adapt them to ensure differential privacy. A salient feature of these algorithms is that they provide optimal algorithms (in terms of error, storage, computation and communication) for locally differentially private heavy hitters.

Outreach: The deployed system currently runs on all iOS and macOS devices. The technology has been used for applications such as learning new words from user keyboards, learning health analytics, and device telemetry.

1.5.3 Case Study: Privacy-preserving Data Mining and Analytics at LinkedIn. We will next highlight the privacy challenges encountered during the design of LinkedIn Salary, a web-scale crowdsourcing system for secure collection and presentation of compensation insights to job seekers [15]. We will describe the privacy mechanisms based on techniques such as encryption, access control, de-identification, aggregation, and thresholding, and present open research challenges in the context of providing rigorous privacy guarantees [16].

We will also present the experiences from deploying differential privacy inspired mechanism for privacy-preserving analytics at LinkedIn [17]. The goal is to compute robust, reliable analytics in a privacy-preserving manner, while satisfying product requirements such as coverage, utility, and consistency. The key idea in this system is to use deterministic pseudorandom noise generation and perform post-processing to achieve data consistency.

1.5.4 Case Study: Microsoft's Local Differential Privacy Deployment for Collecting Windows Telemetry. We will briefly present the mechanisms underlying Microsoft's local differential privacy implementation, which have been deployed across millions of Windows devices to collect application usage statistics in a privacy-preserving manner [7].

1.6 Emerging Topics and Research Directions

In the last part of the tutorial, we will discuss emerging topics pertaining to privacy-preserving data mining and machine learning such as distributed private machine learning [5, 21–23], encode-shuffle-analyze architecture [4], and privacy amplification [13]. Finally, we will discuss open challenges and research directions for the community.

REFERENCES

[1] Learning with privacy at scale. *Apple's Machine Learning Journal*, 1(8), 2017.

[2] M. Barbaro, T. Zeller, and S. Hansell. A face is exposed for AOL searcher no. 4417749. *New York Times*, August 2006.

[3] R. Bassily, K. Nissim, U. Stemmer, and A. Thakurta. Practical locally private heavy hitters. In *NIPS*, 2017.

[4] A. Bittau, U. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld. PROCHLO: Strong privacy for analytics in the crowd. In *SOSP*, 2017.

[5] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth. Practical secure aggregation for privacy-preserving machine learning. In *CCS*, 2017.

[6] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov. "you might also like:" privacy risks of collaborative filtering. In *IEEE Symposium on Security and Privacy*, 2011.

[7] B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In *NIPS*, 2017.

[8] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[10] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 2014.

[11] Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *CCS*, 2014.

[12] G. Fanti, V. Pihur, and Ú. Erlingsson. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, 3, 2016.

[13] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta. Privacy amplification by iteration. In *FOCS*, 2018.

[14] S. Haney, A. Machanavajjhala, J. M. Abowd, M. Graham, M. Kutzbach, and L. Vilhuber. Utility cost of formal privacy for releasing national employer-employee statistics. In *SIGMOD*, 2017.

[15] K. Kenthapadi, S. Ambler, L. Zhang, and D. Agarwal. Bringing salary transparency to the world: Computing robust compensation insights via LinkedIn Salary. In *CIKM*, 2017. Available at https://arxiv.org/abs/1703.09845.

[16] K. Kenthapadi, A. Chudhary, and S. Ambler. LinkedIn Salary: A system for secure collection and presentation of structured compensation insights to job seekers. In *IEEE PAC*, 2017. Available at https://arxiv.org/abs/1705.06976.

[17] K. Kenthapadi and T. T. L. Tran. PriPeARL: A framework for privacy-preserving analytics at LinkedIn. In *CIKM*, 2018.

[18] A. Korolova. Privacy violations using microtargeted ads: A case study. *J. Privacy and Confidentiality*, 3(1), 2011.

[19] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *ICDE*, 2008.

[20] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, 2008.

[21] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*, 2017.

[22] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. Scalable private learning with PATE. In *ICLR*, 2018.

[23] A. Smith, A. Thakurta, and J. Upadhyay. Is interaction necessary for distributed private learning? In *IEEE Symposium on Security and Privacy*, 2017.

[24] J. Su, A. Shukla, S. Goel, and A. Narayanan. De-anonymizing web browsing data with social networks, 2017.

[25] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05), 2002.

[26] P. Voigt and A. von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide.* Springer, 2017.