Differential Privacy for Information Retrieval

ICTIR 2017 Tutorial

Grace Hui Yang
Georgetown University
Washington, D. C.
huiyang@cs.georgetown.edu

Sicong Zhang
Georgetown University
Washington, D. C.
sz303@georgetown.edu

ABSTRACT

Information Retrieval (IR) research has extensively utilized personalization to advance its state-of-the-art. In this process, many IR algorithms and applications require the use of users' personal information, contextual information and other sensitive and private information. However, while IR researchers are making progress, there is always a concern over violations to the users' privacy. Sometimes, the concern becomes so overwhelming that IR research has to stop to avoid leaking users' privacy. The good news is that there have been increasing attentions paid on the joint field of privacy and IR – privacy-preserving IR. As part of the effort, this tutorial offers an introduction to differential privacy (DP), one of the most advanced techniques in privacy research, and provides necessary set of theoretical knowledge for applying privacy techniques in IR. Differential privacy is a technique that provides strong privacy guarantees for data protection. Theoretically, it aims to maximize the data utility in statistical datasets while minimizing the risk of exposing individual data entries to any adversary. Differential privacy has been applied across a wide range of applications in database, data mining, and IR. This tutorial aims to lay a theoretical foundation of DP and how it can be applied to IR. We hope the attendees of this tutorial will have a good understanding of DP and the necessary knowledge to work on this newly minted joint research field of privacy and IR.

KEYWORDS

Differential Privacy; Privacy-Preserving Information Retrieval

1 MOTIVATION

The rapid development of big data, social networks, mobile services and the growing popularity of digital communications have profoundly changed Information Retrieval (IR). Many recent advances in IR research rely on sensitive and private data such as large-scale query logs, users' search history, and location information. It is understandable that the sensitive and private data is kept within commercial companies without being shared with the research community. However, the concern of privacy sometimes is so overwhelming that it has hurt IR research in the past. For instance, the TREC Medical Record Retrieval Tracks [21] are halted because of the privacy issue and the TREC Microblog Tracks [10] could not provide participants with a standard testbed of tweets to ensure a fair comparison. The proper use of privacy techniques to empower privacy-preserving IR [25] should be studied at a timely manner.

The major concerns about privacy in IR include how to properly use personalized data for IR research and how to preserve privacy when releasing them. For instance, web query logs and medical records could not be shared with the public or the researchers without proper treatment. However, without having enough education in privacy research, it is very challenging to study how to make sure data used in IR research can be shared with a certain degree of privacy guarantee and at the same time its IR utility is preserved. This tutorial focuses on the latest technology of differential privacy in particular and how it can be used in IR.

Differential privacy (DP) is the state-of-the-art approach which provides a strong privacy notion and has been widely used in the database and data mining communities. Differential privacy provides guarantees which can be theoretically proved that no individual user in the datasets could be identified. Recent research has shown that differential privacy provides the strongest privacy guarantees among all other privacy techniques and has been shown to be effective in supporting multiple IR tasks [4, 6, 7, 19, 26].

Since privacy-preserving IR is a joint field in both privacy and IR, the success of this field requires researchers from both sides to understand techniques from each other. Therefore, it is necessary to introduce promising privacy techniques such as DP to IR researchers and practitioners. In this tutorial, we focus on introducing the theory of differential privacy as well as how it can be applied to IR research. We cover successful examples of using DP to support IR tasks such as web search, query suggestion, and geological information retrieval. We hope that this tutorial could be a milestone in the development of privacy-preserving IR and enable more valuable research in this promising new joint field.

2 LEARNING OBJECTIVES

The objective of this tutorial is to provide a comprehensive and up-to-date introduction to differential privacy for IR research. We also present a handful of recent IR and mining applications utilizing DP. By the end of this tutorial, the attendees are able to:

- Master DP's mathematical foundation.
- Have a sound understanding of how DP connects to IR.
- Have knowledge of how DP is used in the state-of-the-art research in IR and data mining.
- Be able to generalize the use of DP in other privacy-preserving IR scenarios.

ICTIR '17, October 1–4, 2017, Amsterdam, Netherlands

ACM ISBN 978-1-4503-4490-6/17/10.
https://doi.org/10.1145/3121050.3121107

3 TUTORIAL ORGANIZATION

Tutorial Length: A half-day (3 hours plus breaks).
Prerequisite: Basic knowledge of IR and a good understanding of probability and statistics.

Theme 1 Privacy-Preserving IR and Early Attempts - 50 mins

(1) Background: Privacy concerns in IR. [27]
(2) Privacy-Preserving Information Retrieval (PPIR): General Methodologies.
(3) Naive privacy techniques [3].
(4) K-Anonymity [16], T-Closeness [9], L-Diversity [11].
(5) Recent research topics in Privacy-Preserving IR [14, 24, 25].

Theme 2 Theory of Differential Privacy - 50 mins

(1) Background knowledge in probability.
(2) Mathematical definition of DP [4].
(3) Characteristics of DP.
(4) Analysis about DP.

Theme 3 IR applications using Differential Privacy - 50 mins

(1) Why differential privacy is applicable to IR.
(2) Query Log Anonymization [6, 7, 26, 28].
(3) Geographic IR [18–20].
(4) Other applications that use DP.

Theme 4 Other applications using Differential Privacy - 30 mins

(1) Differential Privacy in Social Network Analysis [17].
(2) Histogram Publication for Dynamic Datasets [8].
(3) Text Mining [5, 22], Frequent Graph Pattern Mining [13] and Data Sequence Mining [1, 2, 15, 23].
(4) Inference from Geo-location Data [12].
(5) Wrapping up the tutorial.

4 CONCLUSIONS

Privacy in IR is an emerging field of research. This tutorial introduces a state-of-the-art privacy technique – differential privacy – to the IR community. The purpose of this tutorial is to provide necessary background knowledge for IR researchers to solve the privacy issues in their related research. Differential privacy is a theoretical framework that requires good mathematical skills and deep understanding to master it. It is not trivial to learn this subject however due to the serious concerns over privacy issues and the strong privacy guarantee provided by this latest technique, we think it is necessary for anyone who would like to pursue research in privacy-preserving IR to master this subject. We hope the tutorial to help lay a solid theoretical foundation for IR researchers and practitioners to use DP to solve many privacy problems in IR.

5 ACKNOWLEDGMENTS

This research was supported by NSF grant IIS-145374 and DARPA grant FA8750-14-2-0226. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors, and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. 2010. Discovering Frequent Patterns in Sensitive Data. In *KDD '10.*
[2] Rui Chen, Gergely Acs, and Claude Castelluccia. 2012. Differentially Private Sequential Data Publication via Variable-length N-grams. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS '12).*
[3] Alissa Cooper. 2008. A Survey of Query Log Privacy-enhancing Techniques from a Policy Perspective. *ACM Trans. Web* 2, 4, Article 19 (Oct. 2008), 27 pages.
[4] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Conference on Theory of Cryptography (TCC'06).*
[5] Arik Friedman and Assaf Schuster. 2010. Data mining with differential privacy. In *KDDâĂŹ10.* ACM, 493–502.
[6] Michaela Gotz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke. 2012. Publishing Search Logs – A Comparative Study of Privacy Guarantees. *IEEE Trans. on Knowl. and Data Eng.* 24, 3 (March 2012).
[7] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing Search Queries and Clicks Privately. In *WWW '09.*
[8] Haoran Li, Li Xiong, Xiaoqian Jiang, and Jinfei Liu. 2015. Differentially Private Histogram Publication for Dynamic Datasets: an Adaptive Sampling Approach. In *CIKM '15.*
[9] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *ICDE 2007.*
[10] Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. 2014. *Overview of the TREC-2014 Microblog track.* Technical Report. DTIC Document.
[11] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L-diversity: Privacy Beyond K-anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1, Article 3 (March 2007).
[12] Cyrus Shahabi, Liyue Fan, Luciano Nocera, Li Xiong, and Ming Li. Privacy-preserving Inference of Social Relationships from Location Data: A Vision Paper. In *SIGSPATIAL '15.* Article 9, 4 pages.
[13] Entong Shen and Ting Yu. 2013. Mining Frequent Graph Patterns with Differential Privacy. In *KDD '13.*
[14] Luo Si, Grace Hui Yang, Sicong Zhang, and Lei Cen. 2014. Proceeding of the 1 st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security. *PIR* (2014).
[15] S. Su, S. Xu, X. Cheng, Z. Li, and F. Yang. 2015. Differentially Private Frequent Itemset Mining via Transaction Splitting. *IEEE Transactions on Knowledge and Data Engineering* 27, 7 (July 2015).
[16] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002).
[17] Christine Task and Chris Clifton. 2012. A Guide to Differential Privacy Theory in Social Network Analysis. In *ASONAM '12.*
[18] Hien To, Liyue Fan, and Cyrus Shahabi. 2015. Differentially Private H-Tree. In *GeoPrivacy'15.* Article 3, 8 pages.
[19] H. To, G. Ghinita, L. Fan, and C. Shahabi. 2017. Differentially Private Location Protection for Worker Datasets in Spatial Crowdsourcing. *IEEE Transactions on Mobile Computing* 16, 4 (April 2017), 934–949.
[20] Hien To, Kien Nguyen, and Cyrus Shahabi. 2016. Differentially Private Publication of Location Entropy. In *GIS '16.* Article 35, 10 pages.
[21] Ellen M Voorhees and William R Hersh. 2012. Overview of the TREC 2012 Medical Records Track.. In *TREC.*
[22] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren. 2014. Information Security in Big Data: Privacy and Data Mining. *IEEE Access* 2 (2014), 1149–1176. DOI: http://dx.doi.org/10.1109/ACCESS.2014.2362522
[23] S. Xu, S. Su, X. Cheng, Z. Li, and L. Xiong. 2015. Differentially private frequent sequence mining via sampling-based candidate pruning. In *ICDE 2015.*
[24] Grace Hui Yang and Ian Soboroff. 2015. Privacy Preserving IR 2015: A SIGIR 2015 Workshop.. In *SIGIR Forum*, Vol. 49. 98–101.
[25] Hui Yang, Ian Soboroff, Li Xiong, Charles L.A. Clarke, and Simson L. Garfinkel. 2016. Privacy-Preserving IR 2016: Differential Privacy, Search, and Social Media. In *SIGIR '16.*
[26] Sicong Zhang, Grace Hui Yang, Lisa Singh, and Li Xiong. 2016. Safelog: Supporting Web Search and Mining by Differentially-Private Query Logs. In *2016 AAAI Fall Symposium Series.*
[27] Sicong Zhang, Hui Yang, and Lisa Singh. 2014. Increased Information Leakage from Text.. In *PIR 2014@ SIGIR.* 41–42.
[28] Sicong Zhang, Hui Yang, and Lisa Singh. 2016. Anonymizing Query Logs by Differential Privacy. In *SIGIR '16.*