Differential Privacy for Information Retrieval

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

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

ABSTRACT

The concern for privacy is real for any research that uses user data. Information Retrieval (IR) is not an exception. Many IR algorithms and applications require the use of users' personal information, contextual information and other sensitive and private information. The extensive use of personalization in IR has become a double-edged sword. Sometimes, the concern becomes so overwhelming that IR research has to stop to avoid privacy leaks. The good news is that recently 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 successfully applied to a wide range of applications in database (DB) and data mining (DM). The research in privacy-preserving IR is relatively new, however, research has shown that DP is also effective in supporting multiple IR tasks. This tutorial aims to lay a theoretical foundation of DP and explains how it can be applied to IR. It highlights the differences in IR tasks and DB and DM tasks and how DP connects to IR. We hope the attendees of this tutorial will have a good understanding of DP and other necessary knowledge to work on the newly minted joint research field of privacy and IR.

ACM Reference Format:
Grace Hui Yang and Sicong Zhang. 2018. Differential Privacy for Information Retrieval. In *WSDM 2018: WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining , February 5–9, 2018, Marina Del Rey, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3159652.3162006

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 became so overwhelming that it hurt IR research in the past. For instance, the TREC Medical Record Retrieval Tracks [21] were halted because of the privacy issue and the TREC Microblog Tracks [10] could not provide participants with a standard testbed of tweets for a fair comparison. The proper use of privacy techniques to empower privacy-preserving IR [25] need to be studied at a timely manner.

Differential privacy (DP) [4] is the state-of-the-art approach which provides a strong privacy notion and has been widely used in the database (DB) and data mining (DM) [2, 5, 8, 15, 17, 22] communities. Recent research has shown that differential privacy provides the strongest privacy guarantees among all other privacy techniques.

Differential privacy is mostly used to protect statistical database, or more generally, for data frequencies. A major challenge of using DP in IR research is the involvement of natural language corpora. Those natural language corpora contain open domains of words, queries, and web documents. Another challenge in applying DP to IR is related to the long-tail effect of the zipf's law. The effect produces very sparse frequency distributions, both at the term level and the document level, which makes many differentially private algorithms in data mining cannot be directly applied to IR research due to very high computational complexity. We will highlight the unique IR challenges in using DP, with comparisons to similar applications in DB and DM.

The major concerns of 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 should not be shared without privacy enhancement. During the recent years, researchers have shown that DP is effective in supporting a few IR topics such as query log anonymization [19] and Geographic IR [13]. In this tutorial, 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 TOPICS TO BE COVERED

Theme 1 Privacy-Preserving IR and Early Attempts - 60 mins

(1) Background: Privacy concerns in IR. [28]
(2) Privacy-Preserving Information Retrieval (PPIR).
(3) Naive privacy techniques [3].
(4) K-Anonymity [16], T-Closeness [9], L-Diversity [12].
(5) Privacy in Search [1, 11] and recent research topics in Privacy-Preserving IR [14, 23, 25].

Theme 2 Differential Privacy - 50 mins

(1) Background knowledge in probability.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02.
https://doi.org/10.1145/3159652.3162006

(2) Mathematical definitions of DP [4].
(3) Discussions about DP.

Theme 3 IR applications using Differential Privacy - 70 mins

(1) Why differential privacy is applicable to IR.
(2) Query Log Anonymization [6, 7, 26, 27, 29].
(3) Geographic IR [13, 18–20].
(4) Other applications that use DP.
(5) Wrap Up and Discussions.

3 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.

4 LINKS TO RELATED RESOURCES

A website to related resources is located at *https://privacypreservingir.org/*. It contains related publications and

- An early version of this tutorial [24] given in the 3rd ACM International Conference on the Theory of Information Retrieval (ICTIR 2017), Amsterdam, Netherlands. Oct 1, 2017.
- The first, second, and third "Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security" Workshops PPIR'14 [14], PPIR'15 [23] and PPIR'16 [25] organized by the authors during the 37-39th International ACM SIGIR Conferences (SIGIR 2014 - 2016).

5 SUPPORT MATERIALS

Attendees will be given printed handouts and a copy of slides.

6 CONCLUSION

Privacy in IR is an emerging field of research. This tutorial highlights the differences of IR applications and other applications in DB and DM and introduces a state-of-the-art privacy technique – differential privacy – to the WSDM community. The purpose of this tutorial is to provide necessary background knowledge to solve the privacy issues in IR 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 serious privacy concerns in IR 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 foundation for using DP to solve many privacy problems in IR.

7 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] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou. 2014. Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data. *IEEE Transactions on Parallel and Distributed Systems* 25, 1 (Jan 2014), 222–233.
[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] W. Lu, A. L. Varna, and M. Wu. 2010. Security analysis for privacy preserving search of multimedia. In *2010 IEEE International Conference on Image Processing*.
[12] 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).
[13] 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.
[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'12*.
[22] 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*.
[23] Grace Hui Yang and Ian Soboroff. 2015. Privacy Preserving IR 2015: A SIGIR 2015 Workshop.. In *SIGIR Forum*, Vol. 49. 98–101.
[24] Grace Hui Yang and Sicong Zhang. 2017. Tutorial: Differential Privacy for Information Retrieval. In *the 3rd ACM International Conference on the Theory of Information Retrieval, ICTIR'17. Amsterdam, Netherlands*.
[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 and Grace Hui Yang. 2017. Deriving Differentially Private Session Logs for Query Suggestion. In *ICTIR'17*.
[27] 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*.
[28] Sicong Zhang, Hui Yang, and Lisa Singh. 2014. Increased Information Leakage from Text.. In *PIR 2014@ SIGIR*. 41–42.
[29] Sicong Zhang, Hui Yang, and Lisa Singh. 2016. Anonymizing Query Logs by Differential Privacy. In *SIGIR '16*.