UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Dottorato di Ricerca in Ingegneria dell'Informazione

Doctoral Course

"Privacy Preserving Information Access"

Prof. Nicola Tonellotto
University of Pisa – Italy
E-mail: nicola.tonellotto@unipi.it

PhD. Guglielmo Faggioli
Università degli Studi di Padova - Italy
E-mail address: guglielmo.faggioli@unipd.it

Short Abstract:

The course aims at allowing PhD students to familiarize with the concept of Privacy, understand its relevance when handling users' data, and learn some practical techniques that can be employed to anonymize and release the data. Privacy protection has ethical [1], legal [2,3] and economic [4] implications that need to be accounted when developing a system which processes, transmits or handles personal and user's generated data. Indeed, privacy plays a prominent role in granting cybersecurity in many digital environments, ranging from customers' data and click logs [5], to medical and genomic data [6,7], to geospatial information [8].

After a brief introduction of the General Data Protection Regulation (GDPR) [2], the major European legal framework that regulates privacy aspects, the course will introduce privacy techniques derived from three major areas: microdata protection, differential privacy, and geomasking.

Microdata are data concerning single individuals. Data used in biological scenarios and generated by sensors typically fall within this category of data. Their use comes with severe risks of reidentification and record linkage [9]. The course will equip the PhD students with statistical techniques to operate securely on such type of data [10]. Furthermore, the course will introduce the main theoretical frameworks to handle this type of information, such as k-Anonymity [11], l-Diversity [12], and t-Closeness [13].

The second module regards Differential Privacy [14]. Differential Privacy is considered the de-facto standard to release privatized data, a paramount task when it comes to digital communications and data publication at large. During the second part of the course, the Phd Students will learn the main basic Differential Privacy mechanisms [14], the building blocks of more advanced solutions, and will be introduced to real world solutions developed by major IT companies, such as Google's RAPPOR [15], Apple's Private CMS [16] and Microsoft's LDP [17].

The final module will concern geographical data. Due to its volume and sensitivity, this class of data presents additional vulnerabilities and requires proper strategies to be handled in a secure manner. To this end, the students will be introduced to the major Geomasking approaches, including statistical solutions [8] and Metric Differential Privacy [18].

Each module will be followed by a hands-on laboratory where the students can learn how to practically implement the techniques discussed theoretically during the lectures. The implementation will be done in Python, using the major data science packages, such as NumPy, Pandas and Scipy. The students will be able to apply privacy protection approaches to the data they use in their research, or on synthetic or publicly available datasets.

Course Contents in brief:

- **Microdata Protection:**
    - Definition of the concept of microdata and related aspects, such as Personal Identifiable Information (PII), Identifiers, Quasi-Identifiers, and sensitive attributes.
    - Analysis of the main micro-data protection techniques, including local suppression, recoding, resampling, Post RAndomized Methods (PRAM), micro-aggregation.
    - Introduction to the main micro-data anonymization frameworks, such as k-Anonymity, l-Diversity, t-Closeness.
- **Differential privacy:**
    - Introduction of the concept of Differential Privacy, with its use cases, application scenarios and limitations.
    - Introduction to the main differentially private mechanisms, including the randomization mechanism, Laplace mechanism, exponential mechanism.
    - Introduction of some real-world applications of Differential Privacy, such as Google's RAPPOR, Apple's Private Count Mean Sketch, and Microsoft's LDP.
- **Biosciences:**
    - Understand the criticalities and additional challenges that rise in terms of privacy when handling medical and biological data, including genomic information.
    - Application of privacy preserving techniques to protect microdata in the medical and biological domains.
    - Devise approaches to produce aggregated statistics on biological data that can be safely released (i.e., published) by employing differential privacy.
- **Automation Engineering:**
    - Identification of the privacy risks derived from handling data generated through sensors.
    - Introduction to the privacy risks associated with geospatial information (reverse geocoding) and main approaches to handle geographical data in a privacy preserving manner (geomasking and metric Differential Privacy).
    - Application of the techniques learned during lecture to protect sensor data and user generated information.
- **Telecommunications:**
    - Analysis of the main approaches studied in the theoretical part of the course explicitly designed for telecommunication data, such as Microsoft's LDP.
    - Application of privacy preserving techniques to telecommunication tasks, such as traffic pattern analysis and resource allocation.
    - Implementation of privacy-preserving solutions specifically tailored towards telecommunication and stream data.

The course will leverage Python for practical implementation of privacy-preserving approaches. Students will be introduced to essential libraries like: NumPy, Pandas, Scipy.

Total # of hours of lecture: 20

References:

[1] United Nations. (1948). Universal Declaration of Human Rights, art 12. https://www.un.org/en/about-us/universal-declaration-of-human-rights

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free

movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)

[3] Artificial Intelligence Act, Regulation (EU) 2024/1689

[4] Cisco. Cisco 2024 Consumer Privacy Survey (2024). https://www.cisco.com/c/en/us/about/trust-center/consumer-privacy-survey.html

[5] Alissa Cooper: A survey of query log privacy-enhancing techniques from a policy perspective. ACM Trans. Web 2(4): 19:1-19:27 (2008)

[6] Bradley A. Malin, Khaled El Emam, Christine M. O'Keefe: Biomedical data privacy: problems, perspectives, and recent advances. J. Am. Medical Informatics Assoc. 20(1): 2-6 (2013)

[7] Mete Akgün, Ali Osman Bayrak, Bugra Ozer, Mahmut Samil Sagiroglu: Privacy preserving processing of genomic data: A survey. J. Biomed. Informatics 56: 103-111 (2015)

[8] Song Gao, Jinmeng Rao, Xinyi Liu, Yuhao Kang, Qunying Huang, Joseph App: Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of Twitter users. J. Spatial Inf. Sci. 19: 105-129 (2019)

[9] Kathleen Benitez, Bradley A. Malin: Evaluating re-identification risks with respect to the HIPAA privacy rule. J. Am. Medical Informatics Assoc. 17(2): 169-177 (2010)

[10] Valentina Ciriani, Sabrina De Capitani di Vimercati, Sara Foresti, Pierangela Samarati: Microdata Protection. Secure Data Management in Decentralized Systems 2007: 291-321

[11] Latanya Sweeney: k-Anonymity: A Model for Protecting Privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst. 10(5): 557-570 (2002)

[12] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, Muthuramakrishnan Venkitasubramaniam: L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1(1): 3 (2007)

[13] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. ICDE 2007: 106-115

[14] Cynthia Dwork, Aaron Roth: The Algorithmic Foundations of Differential Privacy. Found. Trends Theor. Comput. Sci. 9(3-4): 211-407 (2014)

[15] Úlfar Erlingsson, Vasyl Pihur, Aleksandra Korolova: RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. CCS 2014: 1054-1067

[16] Apple Differential Privacy Team. Learning with privacy at scale. Apple Mach. Learn. J, 1(8):1–25, 2017.

[17] Bolin Ding, Janardhan Kulkarni, Sergey Yekhanin: Collecting Telemetry Data Privately. NIPS 2017: 3571-3580

[18] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, Catuscia Palamidessi: Broadening the Scope of Differential Privacy Using Metrics. Privacy Enhancing Technologies 2013: 82-102

CV of the Teacher

Nicola Tonellotto is associate professor at the Department of Information Engineering of the University of Pisa. From 2002 to 2019 he was researcher at the Information Science and Technologies Institute “A. Faedo” of the National Research Council of Italy. His main research interests include Cloud Computing, Web Search, and Information Retrieval, with a particular focus on efficient data processing and neural information retrieval. He co-authored more than 100 papers on these topics in peer reviewed international journals and conferences. He is honorary research fellow in the College of Science & Engineering of the School of Computing Science of the University of Glasgow since 2020 and distinguished member of the ACM since 2023.

Guglielmo Faggioli is a postdoctoral researcher at the University of Padova, Italy. He completed his PhD in 2023 with a thesis on Modelling and Explaining IR System Performance Towards Predictive Evaluation. His research includes Information Retrieval, Evaluation and Performance Prediction, Privacy-Preserving Information Retrieval and Privacy-Preserving Data Analysis. Since 2022, he is Teacher of the Privacy Preserving Information Access course for the master’s degree in Cybersecurity at the University of Padua. His teaching experience also include two years as a Teaching Assistant for the Foundations of Databases (A. Y. 2021-2022) and Web Applications (A.Y. 2020-2021 and 2021-2022) courses in the master courses of Computer Engineering and ICT for Internet and Multimedia at the University of Padova.

Final Exam:

- Presentation of the approaches implemented during the course labs. During the labs at the end of each module, the students will prepare a portfolio analyzing the effectiveness of the techniques they implemented. At the end of the course, the students will be required to present such a portfolio and discuss the privacy risks associated with the data they treated, and the applicability and effectiveness of the different techniques studied during the lectures.
- multiple choice quiz on the content of the lectures. The quiz will be 1h long and will consist of 15 multiple choice questions on the theoretical topics discussed during the lectures.

Room and Schedule

Room: *Aula Riunioni del Dipartimento di Ingegneria dell’Informazione, Via G. Caruso 16, Pisa – Ground Floor*

Schedule:

12/5/2025 – 14:30-18:30

13/5/2025 – 9-13

14/5/2025 – 9-13

15/5/2025 – 9-13

16/5/2025 – 9-13