CAS751: IT Metods in Trustworthy ML November 6, 2024

Final Project Guidelines

Instructor: Shahab Asoodeh

The final projects are meant to give you the opportunity to further explore an aspect of differential privacy and/or fairness that interests you, give you the experience of formulating, carrying out, and presenting an interesting, short-term independent project that is similar to an experience you might have in a career as an applied data scientist or a ML researcher confronting privacy and discrimination issues.

Projects can be done individually or in pairs (for more ambitious projects). Many different types of projects are possible:

- Implement and experimentally evaluate differentially private algorithms (or attacks thereto) on real-life datasets. For example, identify a dataset that resembles a sensitive data use case and a type of statistical analysis that would be useful on such a dataset, implement and tune a differentially private algorithm for that analysis and evaluate the privacy-utility trade-off.
- Critically evaluate an existing algorithm for privacy protection, identify potential vulnerabilities and propose or demonstrate improvements using techniques from the course.
- Explore how the "noisy" results from differential privacy can be properly incorporated into usable scientific or consumer data products in a specific use case.
- Explore how differential privacy might be incorporated into a larger system design (with some particular application domain in mind).
- Seek new theoretical results on some aspect of differential privacy or related topics (especially the nascent field of differentially private fair ML).
- Connect differential privacy to some other area of interest to you (e.g., in CS, signal processing, game theory, economics, network science, finance, sociology, etc).

The steps for various aspects of the project are as follows:

1. **Topic Ideas**: Each student is expected to have a list of 2-3 potential project topics by **November 20** and then discuss them with me. For each topic, you should include the general kind of problem or use case that you would like to address in your project and the general methodology (e.g., is it a theory project or an experimental project or etc). The point of these topic ideas is both to get you thinking about the final project as early as possible and to enable me to give you early feedback and suggestions. I do not expect you to know any details of the project (e.g., methodology, techniques, expected results) at this stage; only big picture of the problem(s) you would like to address.
2. **Presentation**: Each student has 25-30 minutes (20 + 5-10 QA) to present their project during the last week of the course. The presentation should motivate the problem you worked on, describe your approach, and present and interpret the results. Be sure to give proper credit to previous work and clearly distinguish what you've done from what's been done before. There is no format for the presentation, but I personally use PowerPoint + LaTeXiT. Feel free to use any format/software that you think fit. You can find some general tips about presentation here:
   - "How to give a technical presentation" by Michael Ernst.
   - "How to give a great research talk" by Simon Peyton-Jones.
   - "How to give a good research talk" by Stephanie Weirich.

3. Each student needs to submit a report (not more than 6 pages) for the final project a week after the last lecture (latest by **December 11**) in the format provided in the course website. The report is meant to give you the "writing a conference paper" experience: you must begin with the introduction including motivation for the research problems and results and literature review, then mathematically and rigorously formulate the problem and delineate the results and compare with the previously known results. If there are numerical experiments in the paper, it'd be great if you can reproduce the experiments (most papers have released their codes).

Some recommendations: This is just a list of papers that I find interesting. You do not have to choose your project from this list.

Theory papers: Papers marked with asterisk are more theory

- Amplifying Membership Exposure via Data Poisoning, NeurIPS 2022.
- Revealing information while preserving privacy, PODS 2003. This paper predates DP and was key in the privacy formulation that led to DP. Read more about the history of "reconstruction attacks" in this blog post.
- Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data, ICLR 2017.
- Smooth Sensitivity and Sampling in Private Data Analysis, STOC 2007.
- Graphical-model based estimation and inference for differential privacy, ICML 2019.
- On the Privacy Risks of Algorithmic Fairness, EuropS&P 2021.
- Deep Learning with Differential Privacy, CCS 2016.
- A Better Bound Gives a Hundred Rounds: Enhanced Privacy Guarantees via f-divergences, JSAIT 2021.
- Privacy Amplification by Iteration*, FOCS 2018.
- Privacy of Noisy Stochastic Gradient Descent: More Iterations without More Privacy Loss*, NeurIPS 2022.
- Differential Privacy Dynamics of Langevin Diffusion and Noisy Gradient Descent*, NeuriPS 2021.
- Resolving the Mixing Time of the Langevin Algorithm to its Stationary Distribution for Log-Concave Sampling*, COLT 2023.
- Extremal Mechanisms for Local Differential Privacy, JMLR 2015.
- RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, CCS 2014. Local DP deployment in Google.
- Amplification by Shuffling: From Local to Central Differential Privacy via Anonymity*
- Equality of Opportunity in Supervised Learning, NeurIPS 2016.
- Fairness Through Awareness, ITCS 2012.
- Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods and Research, 2018.

- Beyond Adult and COMPAS: Fair Multi-Class Prediction via Information Projection*, NeurIPS 2022.
- From Noisy Fixed-Point Iterations to Private ADMM for Centralized and Federated Learning*, ICML 2023.
- Exactly Minimax-Optimal Locally Differentially Private Sampling*, NeurIPS 2024.
- Locally Private Histograms in All Privacy Regimes, 2024.
- Contraction of Locally Differentially Private Mechanisms, JSAIT 2024.
- Simple Binary Hypothesis Testing under Local Differential Privacy and Communication Constraints, COLT 2023.

Not-to-much theory papers:

- Membership Inference Attacks Against Machine Learning Models, S&P 2017.
- Privacy-Preserving Deep Learning, CCS 2015.
- Winning the NIST Contest: A scalable and general approach to differentially private synthetic data , Winner of 2018 NIST differential privacy synthetic data competition.
- A Central Limit Theorem for Differentially Private Query Answering, NeurIPS 2021.
- Differential privacy for the US Census Bureau. This can be a rather exploratory project. Should you choose this topic, you are expected to present and write an exposition of the ongoing lawsuit: The State of Alabama versus United States Department of Commerce. Starting point can be this blog post. A Survey and Datasheet Repository of Publicly Available US Criminal Justice Datasets, NeurIPS 2022.
- Reconstruction attacks in practice. Should you choose this topic, you are expected to present and provide some interesting examples of reconstruction attacks in practice, as outlined in this blog post.
- Attacks on Deidentification's Defenses. This is a very interesting attack on k-anonymity and was chosen for the distinguished paper award at USENIXSecurity 2022.
- Collecting Telemetry Data Privately, NeurIPS 2017. Deployment of local DP in Microsoft.
- Differential Privacy Team. Deployment of local DP in Apple.
- LinkedIn's Audience Engagements API: A Privacy Preserving Data Analytics System at Scale. Deployment of DP in LinkedIn.
- Machine Bias. Code review: `github.com/probublica/compas-analysis` and `github.com/adebayoj/fairml`.
- Algorithmic decision making and the cost of fairness, KDD 2017.
- A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity, ACM FAT*, 2019.
- Differential Privacy Has Disparate Impact on Model Accuracy, NeurIPS 2019.
- Disparate Impact in Differential Privacy from Gradient Misalignment, ICLR 2023.