Differential privacy for the continual release of statistics — Application to COVID-19 —

- **Keywords:** differential privacy, continual release of statistics, machine learning, COVID
- **Duration:** 6 months
- **Supervisors:**
  Jamal Atif `jamal.atif@lamsade.dauphine.fr`
  Olivier Cappé `olivier.cappe@ens.fr`
  Aurelien Bellet `aurelien.bellet@inria.fr`
- **Place:** MILES Team, LAMSADE, Université Paris Dauphine—PSL
  DI-ENS, Ecole Normale Supérieure–PSL
  Magnet Team, Inria, Cristal, Lille

Context

As we have unfortunately witnessed for almost four years now, monitoring the COVID-19 outbreak requires access to accurate and reliable data at very high frequencies, if not in real time. Indeed, the release of statistics about COVID-19 is critical for informing public health decisions and policies, tracking the spread of the virus at various geographical scales, understanding its impact on hospital activity and quality of care, and enhancing transparency, accountability, and trust towards the general public.

However, the collection and release of such data can pose a significant risk to individuals' privacy, among which re-identification, loss of confidentiality, discrimination, and misuse of data. These risks are even more pronounced when data is released on a continual basis, as malicious actors may be able to exploit the temporal nature of the data to gain effectiveness. To prevent such privacy risks in a robust manner, the current gold standard is to rely on the notion of differential privacy [10]. The work on continual release of statistics with differential privacy guarantees dates back to the early 2010s, with the so-called "tree aggregation" method simultaneously introduced by Dwork et al. [1] and Chan et al. [2]. This technique allows for the release of T running sums (counts) with a privacy loss of $O(logT)$, as opposed to the much larger $O(\sqrt{T}$ loss if one uses a naive composition approach. More recently, there has been renewed interest in tree aggregation in the context of privacy-preserving online learning, see for instance the DP-FTRL algorithm [3] as an alternative to DP-SGD (which achieves the same privacy-utility trade-off without resorting to amplification through sub-sampling). Additionally, recent work has further improved and generalized tree aggregation using a matrix factorization interpretation [4]. However, these results are mainly focused on the continual release of a single simple statistic (typically a running sum). There is still a lack of research on many important problems in private continual release, such as (i) studying other types of statistics and understanding the associated privacy loss, and (ii) considering multiple correlated statistics that involve the same set of individuals, such as those with a hierarchical or graph structure (e.g. hospital admissions by regions, mobility data, contact tracing data, etc). Of course, these challenges are quite generic and our longer term objective is to have a concrete impact on privacy-preserving monitoring of a larger panel of diseases or epidemics, as well as of other indicators related to hospital activity and quality of care. We do however focus on the case of the COVID-19 outbreak to take advantage of the large amount of available data related to the pandemic.

Goals

The primary goal is to develop techniques for enhancing the reliability and effectiveness of differential privacy for the continual release of statistics about COVID-19. his will involve developing and analyzing algorithms and techniques for preserving privacy in the release of COVID-19 statistics, as well as conducting real-world evaluations of differential privacy implementations.

We will build on our work on analyzing COVID-19 data during the second lockdown [5]. In this work, we investigated the relationship between population movement and the spread of the virus, and how to incorporate movement data into our models of the pandemic dynamics. To do so, we relied on movement data provided by Facebook (now Meta) through the "Data for Good" program and on hospital admissions data provided by public authorities.

Organisation

- Gain an understanding of differential privacy framework.
- Survey the literature on existing approaches for the differential privacy of continual release of statistics.
- Develop some of state of the art algorithms and apply them on COVID data.
- Develop and implement new algorithms.
- Publish the results in the form of a research paper at top machine learning conferences (ICML, NeurIPS, ICLR, AISTATS, AAAI) and publish the code on GitHub.

Profile

The ideal candidate will meet the following criteria.

- Pursuing a Masters degree or equivalent in Computer Science, Mathematics, Data Science or Artificial Intelligence
- Mathematical maturity, and a strong theoretical background in probability theory, statistics and machine learning
- Experience in programming (Python) and DL frameworks (PyTorch)
- Exposure to differential privacy and robust machine learning is a plus

References

[1] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, "Differential privacy under continual observation," in *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 715–724, 2010.

[2] T.-H. H. Chan, E. Shi, and D. Song, "Private and continual release of statistics," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 3, pp. 1–24, 2011.

[3] P. Kairouz, B. McMahan, S. Song, O. Thakkar, A. Thakurta, and Z. Xu, "Practical and private (deep) learning without sampling or shuffling," in *International Conference on Machine Learning*, pp. 5213–5225, PMLR, 2021.

[4] C. A. Choquette-Choo, H. B. McMahan, K. Rush, and A. Thakurta, "Multi-epoch matrix factorization mechanisms for private machine learning," *arXiv preprint arXiv:2211.06530*, 2022.

[5] J. Atif, B. Cabot, O. Cappé, O. Mula, and R. Pinot, *Initiative face au virus. Regards croisés sur l'épidemie de Covid-19 apportés par les données sanitaires et de géolocalisation (mars à octobre 2020).* PhD thesis, Université PSL; Inria; CNRS, 2020.