Differential Privacy for Epidemic Surveillance

– **Keywords:** Differential privacy, machine learning, time series, forecasting
– **Duration:** 5 or 6 months

– **Supervision:** Muni Sreenivas Pydi, Jamal Atif (MILES Team, LAMSADE, Université Paris Dauphine—PSL), Olivier Cappé (CSD, DI-ENS, Ecole Normale Supérieure—PSL)[1]

– **Location:** Paris Santé Campus, 75015 Paris
– **Follow up:** Priority will be given to candidates interested by pursuing a PhD thesis on the same topic (fully funded 3 years PhD position available)

Context

Over the past few years, it has become evident that monitoring pandemics like the COVID-19 outbreak necessitates timely access to precise and reliable data, ideally in real-time. Indeed, timely release of accurate statistics is crucial for informing public health decisions, monitoring virus spread, understanding its impact on hospitals, and fostering transparency and trust in the public. However, the collection and release of such data can pose a significant risk to individuals' privacy, including the threats of 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.

The current gold standard approach to prevent such privacy risks is differential privacy [8]. 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 $O(\log T)$, 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], and tight error bounds have been proposed for continual sum/count type queries [7, 6]. However, these results are mainly focused on the continual release of a single simple statistic (typically a running sum).

The challenge of private continual release of data can be further exacerbated because of the constraints imposed by the hierarchical nature of location data. For example, the private counts at different levels of granularity (hospital-level, city-level, region-level etc.) should be consistent across all levels [9]. In addition, defining privacy for complex data structures such as location trajectories

[1]Contact: `muni.pydi@lamsade.dauphine.fr`, `jamal.atif@lamsade.dauphine.fr`, `olivier.cappe@cnrs.fr`

of users of a contact-tracing app (which can be used to track the efficacy of lockdown measures, for example), can pose additional challenges [5, 10]. 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.

Organisation

- Survey the literature on differential privacy of continual release of statistics [2, 1].
- Understand the literature on existing approaches for the differential privacy of hierarchical data [9] and location trajectory data [5, 10].
- Develop new differentially private algorithms for different data modalities like running sums, growth rates estimates, mobility tracking or hierarchical location data, that are suitable for analyzing data collected during the COVID-19 pandemic.

Profile of Candidate

- Pursuing a Masters degree in Computer Science or Mathematics
- Strong theoretical background in probability theory, statistics and machine learning
- Exposure to differential privacy in the form of a masters-level course 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] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1298–1309, 2015.

[6] H. Fichtenberger, M. Henzinger, and J. Upadhyay, "Constant matters: Fine-grained error bound on differentially private continual observation," in *Proceedings of the 40th International Conference on Machine Learning*, pp. 10072–10092, PMLR, 2023.

[7] M. Henzinger, J. Upadhyay, and S. Upadhyay, "Almost tight error bounds on differentially private continual counting," in *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 5003–5039, SIAM, 2023.

[8] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[9] F. Fioretto, P. Van Hentenryck, and K. Zhu, "Differential privacy of hierarchical census data: An optimization approach," *Artificial Intelligence*, vol. 296, p. 103475, 2021.

[10] M. Fiore, P. Katsikouli, E. Zavou, M. Cunche, F. Fessant, D. Le Hello, U. M. Aivodji, B. Olivier, T. Quertier, and R. Stanica, "Privacy in trajectory micro-data publishing: a survey," *Transactions on Data Privacy*, vol. 13, pp. 91–149, 2020.