Differential Privacy for Epidemic Surveillance

Keywords: Differential privacy, machine learning, time series, forecasting

Duration: 5 or 6 months

Supervision: Munni Sreenivas Pydi, Jamal Atif (MILES Team, LAMSADE, Université Paris Dauphine—PSL), Olivier Cappé (CSD, DI-ENS, Ecole Normale Supérieure—PSL)

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

\[ Contact: \text{munni.pydi@lamsade.dauphine.fr, jamal.atif@lamsade.dauphine.fr, olivier.cappe@cnrs.fr} \]
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