Technical Perspective on 'R2T: Instance-optimal Truncation for Differentially Private Query Evaluation with Foreign Keys'

Graham Cormode
University of Warwick, UK
G.Cormode@warwick.ac.uk

Increased use of data to inform decision making has brought with it a rising awareness of the importance of privacy, and the need for appropriate mitigations to be put in place to protect the interests of individuals whose data is being processed. From the demographic statistics that are produced by national censuses, to the complex predictive models built by "big tech" companies, data is the fuel that powers these applications. A majority of such uses rely on data that is derived from the properties and actions of individual people. This data is therefore considered sensitive, and in need of protections to prevent inappropriate use or disclosure. Some protections come from enforcing policies, access control, and contractual agreements. But in addition, we also seek technical interventions: definitions and algorithms that can be applied by computer systems in order to protect the private information while still enabling the intended use.

Although there is not universal consensus, the model of *differential privacy* has emerged as the prevailing notion of privacy, with many deployments in industry and government, and thousands of research papers studying different aspects of the definition [2]. At its heart, different privacy places a requirement on algorithms to introduce uncertainty to their output via randomization, so that the uncertainty is sufficient to provide reasonable doubt over whether the data of any particular person was part of the algorithm's input.

Differential privacy (DP) has proven to be a durable paradigm. First, the definition has been quite robust to attempts to break or circumvent it, whilst being very amenable to generalization to capture different models of privacy, yielding notions where the definition is applied locally to the data of a single user, or enhanced with background knowledge. Second, the definition can be achieved by a wide variety of algorithms applying to different types of data. In particular, there are baseline algorithms that can provide differential privacy for any query with a numeric output: simply compute the true answer and add random noise sampled from a specific statistical distribution (e.g., the Gaussian or Laplace distribution), scaled by the maximum influence that any individual can have on the output. Other generic recipes for privacy involve combining sampling, aggregation, and gradient descent techniques [2, 1].

To put privacy into practice at scale, we need to go beyond developing bespoke algorithms for each particular instance. Instead, we seek to build systems that can handle a broad class of queries specified in a high-level language, and automatically introduce the necessary random noise into the output in order to provide a DP guarantee. In other words, we want a DP-DBMS: a differentially private data management system. The first attempts in this direction were built on top of one of the above-mentioned recipes: PINQ made use of the ability to compute aggregations and add scaled noise [3]; GUPT used the sample-and-aggregate methodology, in conjunction with clipping and noise addition [4].

The limiting factor in deploying differential privacy is often the noise required: we can always achieve DP for some computation, but the noise required may be so large as to render the results meaningless. Standard database queries involving joins are a prime example: the influence of one individual can rapidly blow-up, entailing a very large volume of noise. The natural remedy is to apply truncation: enforcing a hard limit on the influence of an individual, so that we only need to add noise proportional to the truncation limit. This creates a new question: how to choose the calibration limit, which itself has an impact on privacy.

The contribution of this paper is to tackle this tricky calibration problem, and hence enable the development of a DP-DBMS. By using tools from optimization, the techniques developed are able to choose an amount of noise to balance privacy and accuracy. They offer results for the broad class of select-project-join-aggregate (SPJA) queries, and give the first results for queries including self-joins. This opens the way for developing systems that allow queries to performed with privacy guarantees, without requiring the user to have any knowledge of privacy algorithms, or to specify parameters. This supports the long term vision of "privacy everywhere, all at once": allowing privacy to be included in data workflows with zero or minimal configuration.

1. REFERENCES

[1] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM CCS*, 2016.

[2] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[3] F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *ACM SIGMOD*, pages 19–30, 2009.

[4] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. E. Culler. GUPT: privacy preserving data analysis made easy. In *ACM SIGMOD*, pages 349–360, 2012.