CCS 2019 Talk Summaries

Sebastian Meiser, Visa Research

November 11, 2019

This document presents a few summaries of talks I attended at CCS 2019. I don't guarantee that the summaries are comprehensive or that they capture all subtleties of the presented matter; if you find any technical inaccuracies, please contact me about them.

1 Pre-Conference Workshop: TPDP

1.1 Encode, Shuffle, and Analyze (ESA) Revisited: Strong Privacy despite High-Epsilon

Speaker: Abhradeep Guha Thakurta, Google Research, UC Santa Cruz

Assume we have a number of devices use an anonymizer and local DP to achieve some central differential privacy. The anonymizer can either use summation or shuffling to achieve this goal. Naturally we focus on the second part here.

What we do is we take the locally DP objects, remove identifiers (if there are any), shuffle them, and release them. We want to start with weak local DP, i.e., with a high epsilon ($\varepsilon > 1$) and still achieve strong central differential privacy: We get a boost of about $\frac{1}{\sqrt{n}}$ for ε.

To this end, we look at three ideas:

- **Attribute fragmenting:** We split one-hot vectors into the separate bits, then shuffle them, and get some utility in $\Theta\left(\sqrt{\frac{\log k}{ne^{\varepsilon_{\text{local}}}}}\right)$. Note that if we have t records and a local $\varepsilon_{\text{local}} = 1$ we get local DP of $t \cdot \varepsilon_{\text{local}}$.

- **Record fragmenting:** I'm not quite sure what exactly happened here; I think the result is that instead of an ok central DP ($\varepsilon = 1.5$) trade-off that comes with a horrible local DP guarantee ($eps \approx 25$), we can have a local DP of $\varepsilon = 1$ and still get central DP with $\varepsilon = 1.5$. This degrades the utility, but Abhradeep assures us that it's not that bad.

- **Crowds:** We can group records to achieve a better local DP / utility trade-off. We split our data into crowds and analyze them. A cute idea here is to add Laplace noise $\text{Lap}\left(\frac{1}{\varepsilon_{\text{shuffle}}}\right)$ (and subtract a large enough constant) to the count of records it has and then drop as many records as required to meet the count. If we still come up with a number higher than the actual count, we have a distinguishing event.

1.2 DPella

Speaker: Elisabet Lobo Vesga

We know how to do queries with DP and how to estimate the accuracy. However, what happens to our accuracy if we want to add and combine the results of several queries?

In comes DPella, a Haskell library, which allows us to keep track the privacy and accuracy of the (combined) queries we ask and to find out the privacy budget used by a program (via symbolic execution). Similarly, we can get an estimate of the accuracy of the program. That sounds pretty interesting.

So far, they only consider the Laplace mechanism, but they are working on integrating the Gauss mechanism as well.

1.3 Private Stochastic Convex Optimization with Optimal Rate

Speaker: Abhradeep Guha Thakurta, Google Research, UC Santa Cruz

When we design a (differentially private) learning algorithm, we have to consider population risk. Here we have a convex loss function, which makes things easier in many ways. The excess population risk is the expected loss on a random sample, when compared with the minimal such loss. $\max\left(\frac{1}{\sqrt{n}}, \frac{\sqrt{d}}{\varepsilon n}\right)$, the first part of which is what we get for the non-private case and it turns out we often still achieve that.

We look at a noisy SGD with a batch size of approximately $\max\left(\sqrt{n}, \sqrt{d}\right)$; the number of rounds we need is about $\min\left(n, \frac{n^2}{d}\right)$.

1.4 Lessons learned from the NIST DP Synthetic Data Competition

Speaker: Ryan McKenna, University of Massachusetts, Amherst

The talk is about differentially private synthetic data, which is very interesting for a number of reasons. If we have private synthetic data, we can use arbitrary mechanisms on them and we can use it however we want without needing to keep track of any budget.

The speaker discusses the competition he partook in. The data was US census data with 98 attributes and 661k individuals. All the data fields was made of integers between 0 and a known maximum. The synthetic data was judged on: accuracy on all 3-way marginals (10^13 queries) and an approximation of high-order conjunctions (such as "how many records have an age in range X and income in range Y?").

The way they approached this was to first compute a range of noisy aggregate statistics and to then use an inference engine to create synthetic data from it.

1. Measure 1-way marginals using the Gaussian mechanism, treat counts below a threshold as zero.

2. Construct a correlation graph between any pairs of attributes. Each attribute is a node in the graph and the edge weights correspond to the correlation between the attributes. They construct this on the provisional dataset that is assumed to be public (we need to sacrifice the privacy of this dataset). We find a maximum spanning tree of the graph (of the provisional data), then measure 2-way marginals in a differentially private way on the actual data for each part of the tree.

The top 4 solutions of the competition were fairly similar, with the winning one (the one presented) being unique in terms of the inference engine used. Their mechanism ran in 30 minutes.

1.5 Full Convergence of the iterative Bayesian update and applications to local differential privacy

Speaker: Catuscia Palamidessi

In the local DP setting, every user can theoretically set their own privacy level and they don't need to be the same over all users. In this talk we focus on the statistical utility of the data and we try to retrieve the original distribution of the data.

Catuscia presents the iterative Bayesian update: they start with any distribution with full support, e.g., uniform, and then update the distribution iteratively. This approach achieves a pretty good approximation of the original distribution; the work presented fixes a bug of this approach, extends it to more mechanisms, and compares the technique to other inversion techniques.