13

Reflections

13.1 Toward practicing privacy

Differential Privacy was designed with internet-scale data sets in mind. Reconstruction attacks along the lines of those in Section 8 can be carried out by a *polynomial time* bounded adversary asking only $O(n)$ queries on databases of size n. When n is on the order of hundreds of millions, and each query requires a linear amount of computation, such an attack is unrealistic, even thought the queries can be parallelized. This observation led to the earliest steps toward differential privacy: If the adversary is restricted to a *sublinear* number of counting queries, then $o(\sqrt{n})$ noise per query — less than the sampling error! — is sufficient for preserving privacy (Corollary 3.21).

To what extent can differential privacy be brought to bear on smaller data sets, or even targeted attacks that isolate a small subset of a much larger database, without destroying statistical utility? First, an analysis may require a number of queries that begins to look something like the size of this smaller set. Second, letting n now denote the size of the smaller set or small database, and letting k be the number of queries, fractional errors on the order of $\sqrt{k}/n$ are harder to ignore when n is small. Third, the $\sqrt{\ln(1/\delta)}/\varepsilon$ factor in the advanced

composition theorem becomes significant. Keeping in mind the reconstruction attacks when noise is $o(\sqrt{n})$, there appears to be little room to maneuver for arbitrary sets of $k \approx n$ low-sensitivity queries.

There are several promising lines of research for addressing these concerns.

The Query Errors Don't Tell the Whole Story. As an example of this phenomenon, consider the problem of linear regression. The input is a collection of labeled data points of the form (x, y), where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, for arbitrary dimension d. The goal is to find $\theta \in \mathbb{R}^d$ that "predicts" y "as well as possible," given x, under the assumption that the relationship is linear. If the goal is simply to "explain" the given data set, differential privacy may well introduce unacceptable error. Certainly the specific algorithm that simply computes

$$\text{argmin}_\theta | \sum_{i=1}^{n} \theta \cdot x_i - y_i|^2$$

and adds appropriately scaled Laplace noise independently to each coordinate of θ may produce a $\tilde{\theta}$ that differs substantially from θ. But if the goal is to learn a predictor that will do well for *future, unseen* inputs (x, y) then a slightly different computation is used to avoid overfitting, and the (possibly large) difference between the private and non-private coefficient vectors does *not* translate into a gap in classification error! A similar phenomenon has been observed in model fitting.

Less Can Be More. Many analyses ask for more than they actually use. Exploitation of this principle is at the heart of Report Noisy Max, where for the accuracy "price" of one measurement we learn one of the largest of many measurements. By asking for "less" (that is, not requiring that all noisy measurements be released, but rather only asking for the largest one), we obtain "more" (better accuracy). A familiar principle in privacy is to *minimize* collection and reporting. Here we see this play out in the realm of what must be *revealed*, rather than what must be used in the computation.

Quit When You are NOT Ahead. This is the philosophy behind Propose-Test-Release, in which we test in a privacy-preserving way

that small noise is sufficient for a particular intended computation on the given data set.

Algorithms with Data-Dependent Accuracy Bounds. This can be viewed as a generalization of Quit When You are Not Ahead. Algorithms with data-dependent accuracy bounds can deliver excellent results on "good" data sets, as in Propose-Test-Release, and the accuracy can degrade gradually as the "goodness" decreases, an improvement over Propose-Test-Release.

Exploit "Nice" Query Sets. When (potentially large) sets of linear queries are presented as a batch it is possible, by analyzing the geometry of the query *matrix* to obtain higher quality answers than would be obtained were the queries answered independently[1].

Further Relaxation of Differential Privacy We have seen that (ϵ, δ)-differential privacy is a meaningful relaxation of differential privacy that can provide substantially improved accuracy bounds. Moreover, such a relaxation can be essential to these improvements. For example, Propose-Test-Release algorithms can only offer (ε, δ)-differential privacy for $\delta > 0$. What about other, but still meaningful, relaxations of differential privacy? *Concentrated Differential Privacy* is such a relaxation that is incomparable to (ε, δ)-differential privacy and that permits better accuracy. Roughly speaking, it ensures that large privacy loss happens with very small probability; for example, for all k the probability of privacy loss $k\varepsilon$ falls exponentially in k^2. In contrast, (ε, δ)-differential privacy is consistent with having *infinite* privacy loss with probability δ; on the other hand, privacy lost 2ε can happen in concentrated differential privacy with constant probability, while in (ε, δ)-differential privacy it will only occur with probability bounded by δ, which we typically take to be cryptographically small.

Why might we feel comfortable with this relaxation? The answer lies in behavior under composition. As an individual's data participate

[1]More accurately, the analysis is of the object $K = AB_1^k$, where A is the query matrix and B_1^k is the k-dimensional L_1 ball; note that K is the feasible region in answer space when the database has one element.

in many databases and many different computations, perhaps the real worry is the combined threat of multiple exposures. This is captured by privacy under composition. Concentrated differential privacy permits better accuracy while yielding the same behavior under composition as (ε, δ) (and $(\varepsilon, 0)$) differential privacy.

Differential privacy also faces a number of cultural challenges. One of the most significant is non-algorithmic thinking. Differential privacy is a property of an algorithm. However, many people who work with data describe their interactions with the data in fundamentally non-algorithmic terms, such as, "First, I *look at* the data." Similarly, data cleaning is often described in non-algorithmic terms. If data are reasonably plentiful, and the analysts are energetic, then the "Raw Data" application of the Subsample and Aggregate methodology described in Example 7.3 suggests a path toward enabling non-algorithmic, interactions by trusted analysts who will follow directions. In general, it seems plausible that on high-dimensional and on internet-scale data sets non-algorithmic interactions will be the exception.

What about ε? In Example 3.7 we applied Theorem 3.20 to conclude that to bound the cumulative lifetime privacy loss at $\varepsilon = 1$ with probability $1 - e^{-32}$, over participation in $10,000$ databases, it is sufficient that each database be $(1/801, 0)$-differentially private. While $k = 10,000$ may be an overestimate, the dependence on k is fairly weak ($\sqrt{k}$), and in the worst case these bounds are tight, ruling out a more relaxed bound than $\varepsilon_0 = 1/801$ for each database *over the lifetime of the database*. This is simply too strict a requirement in practice.

Perhaps we can ask a different question: Fix ε, say, $\varepsilon = 1$ or $\varepsilon = 1/10$; now ask: How can multiple ε's be apportioned? Permitting ε privacy loss *per query* is too weak, and ε loss over the lifetime of the database is too strong. Something in between, say, ε per study or ε per researcher, may make sense, although this raises the questions of who is a "researcher" and what constitutes a "study." This affords substantially more protection against accidental and intentional privacy compromise than do current practices, from enclaves to confidentiality contracts.

A different proposal is less prescriptive. This proposal draws from second-generation regulatory approaches to reducing environmental

degradation, in particular pollution release registries such as the Toxic Release Inventory that have been found to encourage better practices through transparency. Perhaps a similar effect could arise with private data analysis: an Epsilon Registry describing data uses, granularity of privacy protection, a "burn rate" of privacy loss per unit time, and a cap on total privacy loss permitted before data are retired, when accompanied with a financial penalty for infinite (or very large) loss, can lead to innovation and competition, deploying the talents and resources of a larger set of researchers and privacy professionals in the search for differentially private algorithms.

13.2 The differential privacy lens

An online etymological dictionary describes the original 18th century meaning of the term of the word "statistics" as "science dealing with data about the condition of a state or community." This resonates with differential privacy in the breach: if the presence or absence of the data of a small number of individuals changes the outcome of an analysis then in some sense the outcome is "about" these few individuals, and is not describing the condition of the community as a whole. Put differently, stability to small perturbations in the data is both the hallmark of differential privacy and the essence of a common conception of the term "statistical." Differential privacy is enabled by stability (Section 7) and ensures stability (by definition). In some sense it forces all queries to be statistical in nature. As stability is also increasingly understood to be a key necessary and sufficient condition for learnability, we observe a tantalizing moral equivalence between learnability, differential privacy, and stability.

With this in mind, it is not surprising that differential privacy is also a means to ends other than privacy, and indeed we saw this with game theory in Section 10. The power of differential privacy comes from its amenability to composition. Just as composition allows us to build complex differentially private algorithms from smaller differentially private building blocks, it provides a programming language for constructing stable algorithms for complex analytical tasks. Consider, for example, the problem of eliciting a set of bidder values, and using them to price

a collection of goods that are for sale. Informally, *Walrasian equilibrium prices* are prices such that every individual can simultaneously purchase their *favorite* bundle of goods *given the prices*, while ensuring that demand exactly equals the supply of each good. It would seem at first blush, then, that simply computing these prices, and assigning each person their favorite bundle of goods given the prices would yield a mechanism in which agents were incentivized to tell the truth about their valuation function — since how could any agent do better than receiving their favorite bundle of goods? However, this argument fails — because in a Walrasian equilibrium, agents receive their favorite bundle of goods *given the prices*, but the prices are computed as a function of the reported valuations, so an industrious but dishonest agent could potentially gain by manipulating the computed prices. However, this problem is solved (and an approximately truthful mechanism results) if the equilibrium prices are computed using a differentially private algorithm — precisely because individual agents have almost no effect on the distribution of prices computed. Note that this application is made possible by the use of the tools of differential privacy, but is completely orthogonal to privacy concerns. More generally, this connection is more fundamental: computing *equilibria* of various sorts using algorithms that have the stability property guaranteed by differential privacy leads to approximately truthful mechanisms implementing these equilibrium outcomes.

Differential privacy also helps in ensuring generalizability in adaptive data analysis. Adaptivity means that the questions asked and hypotheses tested depend on outcomes of earlier questions. Generalizability means that the outcome of a computation or a test on the data set is close to the ground truth of the distribution from which the data are sampled. It is known that the naive paradigm of answering queries with the exact empirical values on a fixed data set fails to generalize even under a limited amount of adaptive questioning. Remarkably, answering with differential privacy not only ensures privacy, but with high probability it ensures generalizability even for exponentially many adaptively chosen queries. Thus, the deliberate introduction of noise using the techniques of differential privacy has profound and promising implications for the validity of traditional scientific inquiry.