Differential Privacy

CS 261

1 Definitions

Let us first look at the formal definitions of differential privacy. First, we will need the notion of adjacent databases.

Definition 1.1 (Adjacent Databases)**.** Let $\mathcal{X}$ be a domain and let $\mathbb{N}^{\mathcal{X}}$ denote the set of databases with entries in $\mathcal{X}$. Two datasets B and B' are said to be adjacent if

$$\sum_{x \in \mathcal{X}} \left| B(x) - B'(x) \right| \leq 1.$$

Intuitively, this captures the notion of two databases differing in one entry. With this in hand, we formally define differential privacy.

Definition 1.2 (Differential Privacy)**.** Let $\mathcal{M}$ be a randomized mechanism with inputs in $\mathbb{N}^{\mathcal{X}}$. Let $\epsilon, \delta \geq 0$. $\mathcal{M}$ is said to be (ϵ, δ)-differentially private if for all S, we have

$$\Pr\left[\mathcal{M}(B) \in S\right] \leq e^{\epsilon} \cdot \Pr\left[\mathcal{M}\left(B'\right) \in S\right] + \delta$$

for any two adjacent databases B, B'. If $\delta = 0$, $\mathcal{M}$ is said to be ϵ purely differentially private.

Let us try to understand the definition a bit better. Firstly, the interesting regime is when $\epsilon \ll 1$ (though it is not necessarily achieved in practice). In this regime, $e^{\epsilon} \approx (1 + \epsilon)$. Thus, the definition is saying that for any event the probability of that event happening for B is a $(1 + \epsilon)$ factor of the same event occurring for an adjacent database B'. One reason for the choice of e^{ϵ} is that it behaves better with respect to composition of guarantees of mechanisms. Another fact to notice is that adjacency is a symmetric relation. Thus, though the definition only refers to an upper bound on the probability, it also implies a lower bound on the probabilities of events for adjacent datasets.

The definition above is a rather clean and satisfactory formalization of the notion of privacy. In particular, it satisfies many desirable qualities such as postprocessing and composition. Also, it turns out that there are rich connections between differential privacy and areas such as adaptive data analysis, generalization, online optimization, convex geometry and more. For a more thorough discussion, look at the excellent introductory chapters in [DR14] or [Vad17].

2 History

In this section, we will briefly look at some historical context for differential privacy. One of the institution that is particular about privacy is the US Census bureau. The data collected by the census is important in many contexts such as apportioning representation. Since census requires people to give person data any loss of privacy might lead to people not responding accurately, and generally to loss of faith in the census. This is compounded by a history of abuse in the past. One early solution is aggregation, where only average information is released. But researchers also want finer grain access to the data. Allowing arbitrary queries is clearly problematic. Maybe a solution is to remove names. There is a issue with this as well if the other field are sufficient to uniquely pin down a person. A another solution is to only allow queries to be returned if there many entries matching the query. This notion is known as k-anonymity. Unfortunately, this notion is also broken. The idea is to construct two queries q_1 and q_2 with large number of matches and for which we know the only element in the difference is the target. From this it is easy reconstruct the answer for the target by taking the difference. Such a pair of queries can be easily

constructed by taking a query in which the target is not present and then taking an or with a query that uniquely identifies the target. More complicated queries can be constructed to circumvent additional modifications to this definition. The common wisdom amongst the database community was that support for complicated queries was essentially impossible, until the formulation of differential privacy in 2006.

Medical community were also concerned about these questions since giving researchers access to medical data could lead to many lives being saved. But this is in opposition to confidentiality. This database of patients' data can be used to identify someone. Issues with this was spectacularly demonstrated by L. Sweeney, who used the records available in the state of Massachusetts to identify the governor of Massachusetts and mailed the medical records to the governor's office.

3 Limitations

- Tradeoff between utility and privacy
- Assumption there is only data entry per person
- Group privacy
- Non independent entries can lead to privacy loss; example location data

4 RAPPOR

Randomized Aggregatable Privacy-Preserving Ordinal Response [EPK14], also known as RAPPOR, is a mechanism designed to aggreagate user data in a differentially private way. First let us look at the randomized response mechanism.

Say in a class the instructor wants to figure out if the students want a final exam. Since the students don't necessarily want their preferences to be known, we would like to incorporate differential privacy in the reporting mechanism. Here is a one mechanism that aims to do this

- For each student
  - Roll a die to get answer X
  - If the answer is in $X \in \{1, 2, 3, 4\}$, respond honestly.
  - Else, respond incorrectly.

One can now use the fraction of people responding to get a noisy estimate of the actual number of people who want the final exam. Privacy is guaranteed by the noisyness of each of the responses. By changing the probability of honest response, one can balance privacy and accuracy.

Now, let us consider a scenario where the above poll is run everyday. The above mechanism run independently has the issue that an adversary observing the responses of any person can reconstruct the person's true response by averaging. One solution that RAPPOR uses is to couple all the noise terms amongst all future responses.

Let us see how to deal with correlated questions. Let $Q_1, \ldots, Q_m$ be m queries. Since Q_i could be correlated, knowing the answers could lead to inference. To deal with this, we briefly introduce the Bloom filter data structure. Pick k hash functions $h_1, \ldots, h_k : [m] \to [n]$, where $n < m$. Instantiate an empty array B of length n. Let $Q_1, \ldots, Q_m$ be queries and $A_1, \ldots, A_m$ be the corresponding answers. For each i such that $A_i = 1$, set $B\left[h_j\left(i\right)\right] = 1$ for all j. The idea from RAPPOR is to encode the answers to $Q_1, \ldots, Q_m$ in a Bloom filter as described above and then flipping each bit in the Bloom filter with some probability. This structure does not leak too much information about a single person while still having enough information to recover aggregate information.

One of the main application of this algorithm is in the chrome browser. A specific example is when the Chrome team wanted to understand the usage of Silverlight, in order to know whether it was ok to turn off the service. Naively reporting domain names of the websites that users visit is very intrusive and is generally frowned upon. Google used the RAPPOR scheme to collect the aggregate data.

5 Applications

- US Census in 2020 (they also made this video in collaboration with Minute physics)
- Telemetry:
  - Google chrome (using RAPPOR as discussed above)
  - Apple with Emoji data
  - Shortcoming: Apple's solution does not deal with data across time
  - Google and Apple have been reported to use ϵ as large as 9. This is essentially meaningless from the privacy point of view. This seems be work better in practice than theory suggests.
- GDPR compliance: The right to be forgotten is a legal requirement in certain jurisdiction. Training with differential privacy provably doesn't depend (much) on any single user's data.
- Federated Learning

References

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 1054–1067, New York, NY, USA, 2014. Association for Computing Machinery.

[Vad17] Salil Vadhan. *The Complexity of Differential Privacy*, pages 347–450. Springer International Publishing, Cham, 2017.