# Make Up Your Mind: The Price of Online Queries in Differential Privacy (Short Abstract)

Mark Bun* Thomas Steinke† Jonathan Ullman‡

April 28, 2016

*Differential privacy* [DMNS06] is a formal guarantee that an algorithm run on a sensitive dataset does not reveal too much about any individual in that dataset. Since its introduction, a rich literature has developed to determine what statistics can be computed accurately subject to differential privacy. For example, suppose we wish to approximate a real-valued *query* $q(x)$ on some dataset $x$ that consists of the private data of many individuals. Then, this question has a clean answer—we can compute a differentially private estimate of $q(x)$ with error proportional to the *global sensitivity* of $q$, and we cannot have smaller error in the worst case.

But how much error do we need to answer a large set of queries $q_1,\dots,q_k$? Before we can answer this question, we have to define a model of how the queries are asked and answered. The literature on differential privacy has considered three different interactive models[1] for specifying the queries:

- *The Offline Model:* The sequence of queries $q_1,\dots,q_k$ are given to the algorithm together in a batch and the mechanism answers them together.
- *The Online Model:* The sequence of queries $q_1,\dots,q_k$ is chosen in advance and then the mechanism must answer each query $q_j$ before seeing $q_{j+1}$.
- *The Adaptive Model:* The queries are not fixed in advance, each query $q_{j+1}$ may depend on the answers to queries $q_1,\dots,q_j$.

In all three cases, we assume that $q_1,\cdots,q_k$ are chosen from some family of allowable queries $Q$, but may be chosen adversarially from this family.

Differential privacy seems well-suited to the adaptive model. Arguably its signature property is that any adaptively-chosen sequence of differentially private algorithms remains collectively differentially private, with a graceful degradation of the privacy parameters [DMNS06, DRV10]. As a consequence, there is a simple differentially private algorithm that takes a dataset of $n$ individuals and answers $\tilde{\Omega}(n)$ statistical queries in the adaptive model with error $o(1/\sqrt{n})$,

*Harvard University John A. Paulson School of Engineering and Applied Sciences.
†Harvard University John A. Paulson School of Engineering and Applied Sciences.
‡Northeastern University College of Computer and Information Science.

[1] Usually, the "interactive model" refers only to what we call the "adaptive model." We prefer to call all of these models interactive, since they each require an interaction with a data analyst who issues the queries. We use the term "interactive" to distinguish these models from one where the algorithm only answers a fixed set of queries.

simply by perturbing each answer independently with carefully calibrated noise. In contrast, the seminal lower bound of Dinur and Nissim and its later refinements [DN03, DY08] shows that there exists a fixed set of $O(n)$ queries that cannot be answered by any differentially private algorithm with such little error, even in the easiest offline model. For an even more surprising example, the private multiplicative weights algorithm of Hardt and Rothblum [HR10] can in many cases answer an exponential number of arbitrary, adaptively-chosen statistical queries with a strong accuracy guarantee, whereas [BUV14] show that the accuracy guarantee of private multiplicative weights is nearly optimal even for a simple, fixed family of queries.

These examples might give the impression that answering adaptively chosen queries comes "for free" in differential privacy—that everything that can be achieved in the offline model can be matched in the adaptive model. Beyond just the lack of any separation between the models, many of the most powerful differentially private algorithms in all of these models use techniques from no-regret learning, which are explicitly designed for adaptive models.

In this work, we show for the first time that these three models are actually distinct. In fact, we show exponential separations between each of the three models. These are the first separations between these models in differential privacy.

## 0.1 Our Results

Given a dataset $x$ whose elements come from a data universe $X$, a *statistical query on* $X$ is defined by a predicate $\phi$ on $X$ and asks "what fraction of elements in the dataset satisfy $\phi$?" The answer to a statistical query lies in $[0,1]$ and our goal is to answer these queries up to some small additive error $\pm\alpha$, for a suitable choice of $0 < \alpha < 1$. If the mechanism is required to answer *arbitrary* statistical queries, then the offline, online, and adaptive models are essentially equivalent — the upper bounds in the adaptive model match the lower bounds in the offline model [HR10, BUV14, SU15]. However, we show that when the predicate $\phi$ is required to take a specific form, then it becomes strictly easier to answer a set of these queries in the offline model than it is to answer a sequence of queries presented online.

**Theorem 0.1** (Informal)**.** *There exists a data universe $X$ and a family of statistical queries $Q$ on $X$ such that for every $n \in \mathbb{N}$,*

1. *there is a differentially private algorithm that takes a dataset $x \in X^n$ and answers any set of $k = 2^{\Omega(\sqrt{n})}$ offline queries from $Q$ up to error $\pm 1/100$ from $Q$, but*
2. *no differentially private algorithm can take a dataset $x \in X^n$ and answer an arbitrary sequence of $k = O(n^2)$ online (but not adaptively-chosen) queries from $Q$ up to error $\pm 1/100$.*

This result establishes that the online model is strictly harder than the offline model. We also demonstrate that the adaptive model is strictly harder than the online model. Here, the family of queries we use in our separation is not a family of statistical queries, but is rather a family of *search queries* with a specific definition of accuracy that we will define later.

**Theorem 0.2** (Informal)**.** *For every $n \in \mathbb{N}$, there is a family of "search" queries $Q$ on datasets in $X^n$ such that*

1. *there is a differentially private algorithm that takes a dataset $x \in \{\pm 1\}^n$ and accurately answers any online (but not adaptively-chosen) sequence of $k = 2^{\Omega(n)}$ queries from $Q$, but*

2. *no differentially private algorithm can take a dataset $x \in \{\pm 1\}^n$ and accurately answer an adaptively-chosen sequence of $k = O(1)$ queries from $Q$.*

We leave it as an interesting open question to separate the online and adaptive models for statistical queries, or to show that the models are equivalent for statistical queries.

Although Theorems 0.1 and 0.2 separate the three models, these results use somewhat contrived families of queries. Thus, we also investigate whether the models are distinct for *natural* families of queries that are of use in practical applications. One very well studied class of queries is *threshold queries*. These are a family of statistical queries $Q_{\text{thresh}}$ defined on the universe $[0,1]$ and each query is specified by a point $\tau \in [0,1]$ and asks "what fraction of the elements of the dataset are at most $\tau$?" If we restrict our attention to so-called pure differential privacy (i.e. $(\varepsilon, \delta)$-differential privacy with $\delta = 0$), then we obtain an exponential separation between the offline and online models for answering threshold queries.

**Theorem 0.3** (Informal)**.** *For every $n \in \mathbb{N}$,*

1. *there is a pure differentially private algorithm that takes a dataset $x \in [0,1]^n$ and answers any set of $k = 2^{\Omega(n)}$ offline queries from $Q_{thresh}$ up to error $\pm 1/100$, but*

2. *no pure differentially private algorithm takes a dataset $x \in [0,1]^n$ and answers an arbitrary sequence of $k = O(n)$ online (but not adaptively-chosen) queries from $Q_{thresh}$ up to error $\pm 1/100$.*

We also ask whether or not such a separation exists for arbitrary differentially private algorithms (i.e. $(\varepsilon, \delta)$-differential privacy with $\delta > 0$). Theorem 0.3 shows that, for pure differential privacy, threshold queries have near-maximal sample complexity. That is, up to constants, the lower bound for online threshold queries matches what is achieved by the Laplace mechanism, which is applicable to arbitrary statistical queries. This may lead one to conjecture that adaptive threshold queries also require near-maximal sample complexity subject to approximate differential privacy. However, we show that this is not the case:

**Theorem 0.4.** *For every $n \in \mathbb{N}$, there is a differentially private algorithm that takes a dataset $x \in [0,1]^n$ and answers any set of $k = 2^{\Omega(n)}$ adaptively-chosen queries from $Q_{thresh}$ up to error $\pm 1/100$.*

In contrast, for any offline set of $k$ thresholds $\tau_1, \ldots, \tau_k$, we can round each element of the dataset up to an element in the finite universe $X = \{\tau_1, \ldots, \tau_k, 1\}$ without changing the answers to any of the queries. Then we can use known algorithms for answering all threshold queries over any finite, totally ordered domain [BNS13, BNSV15] to answer the queries using a very small dataset of size $n = 2^{O(\log^*(k))}$. We leave it as an interesting open question to settle the complexity of answering adaptively-chosen threshold queries in the adaptive model.

## References


[BNS13] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*, pages 363–378, 2013.

[BNSV15] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil P. Vadhan. Differentially private release and learning of threshold functions. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 634–649, 2015.

[BUV14] Mark Bun, Jonathan Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 1–10, 2014.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, pages 265–284, 2006.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210, 2003.

[DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *IEEE Symposium on Foundations of Computer Science (FOCS '10)*, pages 51–60. IEEE, 23–26 October 2010.

[DY08] Cynthia Dwork and Sergey Yekhanin. New efficient attacks on statistical disclosure control mechanisms. In *Advances in Cryptology - CRYPTO 2008, 28th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2008. Proceedings*, pages 469–480, 2008.

[HR10] Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 61–70, 2010.

[SU15] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *CoRR*, abs/1501.06095, 2015.