*A*PEx: Accuracy-Aware Differentially Private Data Exploration

Chang Ge
University of Waterloo
c4ge@uwaterloo.ca

Xi He
University of Waterloo
xi.he@uwaterloo.ca

Ihab F. Ilyas
University of Waterloo
ilyas@uwaterloo.ca

Ashwin Machanavajjhala
Duke University
ashwin@cs.duke.edu

ABSTRACT

Organizations are increasingly interested in allowing external data scientists to explore their sensitive datasets. Due to the popularity of differential privacy, data owners want the data exploration to ensure provable privacy guarantees. However, current systems for answering queries with differential privacy place an inordinate burden on the data analysts to understand differential privacy, manage their privacy budget, and even implement new algorithms for noisy query answering. Moreover, current systems do not provide any guarantees to the data analyst on the quality they care about, namely accuracy of query answers.

We present *A*PEx, a novel system that allows data analysts to pose adaptively chosen sequences of queries along with required *accuracy bounds*. By translating queries and accuracy bounds into differentially private algorithms with the least privacy loss, *A*PEx returns query answers to the data analyst that meet the accuracy bounds, and proves to the data owner that the entire data exploration process is differentially private. Our comprehensive experimental study on real datasets demonstrates that *A*PEx can answer a variety of queries accurately with moderate to small privacy loss, and can support data exploration for entity resolution with high accuracy under reasonable privacy settings.

KEYWORDS

Differential Privacy; Data Exploration

SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5643-5/19/06…$15.00
https://doi.org/10.1145/3299869.3300092

ACM Reference Format:
Chang Ge, Xi He, Ihab F. Ilyas, and Ashwin Machanavajjhala. 2019. *A*PEx: Accuracy-Aware Differentially Private Data Exploration. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3299869.3300092

1 INTRODUCTION

Data exploration has been gaining more importance as large and complex datasets are collected by organizations to enable data scientists to augment and to enrich their analysis. These datasets are usually a mix of public data and private sensitive data, and exploring them involves operations such as profiling, summarization, building histograms on various attributes, and top-k queries to better understand the underlying semantics. Besides the large body of work on interactive exploration on large data repositories [4], interacting with large datasets can be an important step towards many mundane jobs. For example, in entity matching tools (such as Magellan [21], and Data Tamer [34]), a sequence of exploration queries are needed to approximate the data distribution of various fields to pick the right matching model parameters, features, etc. In data integration systems, exploring the actual data instance of the sources plays a crucial role in matching the different schemas.

In our interaction with many large enterprises, exploring private data sets to enable large-scale data integration and analytics projects is often a challenging task; while incorporating these data sources in the analysis carries a high value, data owners often do not trust the analysts and hence, require privacy guarantees. For instance, Facebook recently announced that they would allow academics to analyze their data to "study the role of social media in elections and democracy", and one of the key concerns outlined was maintaining the privacy of their customers' data [19, 20].

Differential privacy [8, 10] has emerged as a popular privacy notion since (1) it is a persuasive mathematical guarantee that individual records are hidden even when aggregate statistics are revealed, (2) it ensures privacy even in the presence of side information about the data, and (3) it allows one to bound the information leakage by a total *privacy budget* across multiple data releases. Differential privacy has seen adoption in a number of real products at the US Census Bureau [14, 30, 35], Google [11], Apple [13] and Uber [18]. Hence, it is natural for data owners to aspire for systems that permit differentially private data exploration.

1.1 The Challenge and The Problem

While there exist general purpose differentially private query answering systems, they are not really meant to support interactive querying, and they fall short in two key respects. First, these systems place an inordinate burden on the data analyst to understand differential privacy and differentially private algorithms. For instance, PINQ [32] and wPINQ [33] allow users to write differentially private programs and ensure that every program expressed satisfies differential privacy. However, to achieve high accuracy, the analyst has to be familiar with the privacy literature to understand how the system adds noise and to identify if the desired accuracy can be achieved in the first place. ϵktelo [37] has high level operators that can be composed to create accurate differentially private programs to answer counting queries. However, the analyst still needs to know how to optimally apportion privacy budgets across different operators. FLEX [18] allows users to answer one SQL query under differential privacy, but has the same issue of apportioning privacy budget across a sequence of queries. Second, and somewhat ironically, these systems do not provide any guarantees to the data analyst on the quality they really care about, namely *accuracy* of query answers. In fact, most of these systems take a privacy level (ϵ) as input and make sure that differential privacy holds, but leave accuracy unconstrained.

We aim to design a system that allows data analysts to explore a sensitive dataset D held by a data owner by posing a sequence of declaratively specified queries that can capture typical data exploration workflows. The system aims at achieving the following dual goals: (1) since the data are sensitive, the data owner would like the system to provably bound the information disclosed about any one record in D to the analyst; and (2) since privacy preserving mechanisms introduce error, the data analyst must be able to specify an accuracy bound on each query.

Hence, our goal is to design a system that can:

- Support declaratively specified aggregate queries that capture a wide variety of data exploration tasks.
- Allow analysts to specify accuracy bounds on queries.
- Translate an analyst's query into a differentially private mechanism with minimal privacy loss ϵ such that it can answer the query while meeting the accuracy bound.
- Prove that for any interactively specified sequence of queries, the analyst's view of the entire data exploration process satisfies B-differential privacy, where B is a owner specified privacy budget.

Our problem is similar in spirit to Ligett et al. [29], which also considers analysts who specify accuracy constraints. While their work focus on finding the smallest privacy cost for a given differentially private mechanism and accuracy bound, our focus is on a more general problem: for a given query, find a mechanism and a minimal privacy cost to achieve the given accuracy bound. We highlight the main technical differences in Appendix D.

1.2 Contributions and Solution Overview

We propose *A*PEx, an accuracy-aware privacy engine for sensitive data exploration. *A*PEx solves the aforementioned challenges as follows:

- A data analyst can interact with the private data through *A*PEx using declaratively specified aggregate queries. Our query language supports three types of aggregate queries: (1) *workload counting queries* that capture the large class of linear counting queries (e.g., histograms and CDFs) which are a staple of statistical analysis, (2) *iceberg queries*, which capture HAVING queries in SQL and frequent pattern queries, and (3) *top-k queries*. These queries form the building blocks of several data exploration workflows. To demonstrate their applicability in real scenarios, in Section 8, we express two important data cleaning tasks, namely blocking and pair-wise matching, using sequences of queries from our language.
- In our language, each query is associated with intuitive accuracy bounds that permit *A*PEx to use differentially private mechanisms that introduce noise while meeting the accuracy bound.
- For each query in a sequence, *A*PEx employs an *accuracy translator* that finds a privacy level ϵ and a differentially private mechanism that answers the query while meeting the specified accuracy bound. For the same privacy level, the mechanism that answers a query with the least error depends on the query and dataset. Hence, *A*PEx implements a suite of differentially private mechanisms for each query type, and given an accuracy bound chooses the mechanism that incurs the least privacy loss based on the input query and dataset.
- *A*PEx uses a *privacy analyzer* to decide whether or not to answer a query such that the privacy loss to the analyst is always bounded by a budget B. The privacy analysis is novel since (1) the privacy loss of each mechanism is chosen based on the query's accuracy requirement, and (2) some mechanisms have a data dependent privacy loss.

- In a comprehensive empirical evaluation on real datasets with query and application benchmarks, we demonstrate that (1) *A*PEx chooses a differentially private mechanism with the least privacy loss that answers an input query under a specified accuracy bound, and (2) allows data analysts to accurately explore data while ensuring provable guarantee of privacy to data owners.

1.3 Organization

Section 2 introduces preliminaries and background. Section 3 describes our query language and accuracy measures. *A*PEx architecture is summarized in Section 4 and details of the accuracy translator and the privacy analysis are described in Sections 5 and 6, respectively. *A*PEx is comprehensively evaluated on real data using exploration queries in Section 7 and on an entity resolution case study in Section 8. We conclude with future work in Section 9.

2 PRELIMINARIES AND BACKGROUND

We consider the sensitive dataset in the form of a single-table relational schema $R(A_1, A_2, \ldots, A_d)$, where $attr(R)$ denotes the set of attributes of R. Each attribute A_i has a domain $dom(A_i)$. The full domain of R is $dom(R) = dom(A_1) \times \cdots \times dom(A_d)$, containing all possible tuples conforming to R. An instance D of relation R is a multiset whose elements are tuples in $dom(R)$. We let the domain of the instances be $\mathcal{D}$. Extending our algorithms to schemas with multiple tables is an interesting avenue for future work.

We use differential privacy as our measure of privacy. An algorithm that takes as input a table D satisfies differential privacy [9, 10] if its output does not significantly change by adding or removing a single tuple in its input.

Definition 2.1 (ϵ-Differential Privacy [10]). A randomized mechanism $M : \mathcal{D} \rightarrow \mathcal{O}$ satisfies ϵ-differential privacy if

$$Pr[M(D) \in O] \leq e^{\epsilon} Pr[M(D') \in O] \quad (1)$$

for any set of outputs $O \subseteq \mathcal{O}$, and any pair of *neighboring* databases D, D' such that $|D \backslash D' \cup D' \backslash D| = 1$.

Smaller values of ϵ result in stronger privacy guarantees as D and D' are harder to distinguish using the output. Composition and postprocessing theorems of differential privacy are described in Appendix B and will be used to bound privacy across multiple data releases in Section 6.

3 QUERIES AND ACCURACY

In this section, we describe our query language for expressing aggregate queries and the associated accuracy measures. Throughout the paper, we assume that the schema and the full domain of attributes are public.

3.1 Exploration Queries

*A*PEx supports a rich class of aggregate queries that can be expressed in a SQL-like declarative format.

BIN D ON $f(\cdot)$ WHERE $W = \{\phi_1, \ldots, \phi_L\}$
[HAVING $f(\cdot) > c$]
[ORDER BY $f(\cdot)$ LIMIT k];

Each query in our language is associated with a *workload* of predicates $W = \{\phi_1, \ldots, \phi_L\}$. Based on W, the tuples in a table D are divided into bins. Each bin b_i contains all the tuples in D that satisfy the corresponding predicate $\phi_i : dom(R) \rightarrow \{0, 1\}$, i.e., $b_i = \{r \in D | \phi(r) = 1\}$. As we will see later, bins need not be disjoint. Moreover, a query has an aggregation function $f : dom(R)^* \rightarrow \mathbb{R}$, which returns a numeric answer $f(b_i)$ for each bin b_i. The output of this query without the optional clauses (in square brackets) is a list of counts $f(b_i)$ for bin b_i.

Each query can be specialized using one of two optional clauses: the HAVING clause returns a list of bin identifiers b_i for which $f(b_i) > c$; and the ORDER BY ... LIMIT clause returns the k bins that have the largest values for $f(b_i)$.

Throughout this paper, we focus on COUNT as the aggregate function and discuss other aggregates like AVG, SUM, QUANTILE in Appendix E.

Workload Counting Query (WCQ).

BIN D ON $f(\cdot)$ WHERE $W = \{\phi_1, \ldots, \phi_L\}$;

Workload counting queries capture the large class of *linear counting queries*, which are the bread and butter of statistical analysis and have been the focus of majority of the work in differential privacy [16, 27]. Standard SELECT...GROUP BY queries in SQL are expressible using WCQ. For instance, consider a table D having an attribute *State* with domain {AL, AK, ..., WI, WY} and an attribute *Age* with domain $[0, \infty)$. Then, a query that returns the number of people with age above 50 for each state can be expressed using WCQ as:

BIN D ON COUNT(*)
WHERE $W = \{Age{>}50 \wedge State{=}\text{AL}, \ldots, Age{>}50 \wedge State{=}\text{WY}\}$;

Other common queries captured by WCQ include: (1) *histogram queries:* the workload W partitions D into $|W_h|$ disjoint bins, e.g. $W_h = \{0 < Age \leq 10, 10 < Age \leq 20, \ldots, 90 < Age\}$ the resulting WCQ returns counts for each bin; (2) *cumulative histograms:* we can define a workload W_p that places tuples in D into a set of inclusive bins $b_1 \subseteq b_2 \cdots \subseteq b_L$, e.g. $W_p = \{Age \leq 10, Age \leq 20, \ldots, Age \leq 90\}$. The resulting query outputs a set of cumulative counts. We call such a W_p a *prefix* workload.

Iceberg Counting Query (ICQ).

BIN D ON COUNT(*) WHERE $W = \{\phi_1, \ldots, \phi_L\}$
HAVING COUNT(*) $> c$;

An iceberg query returns bin identifiers if the aggregate value for that bin is greater than a given threshold c. For instance, a query which returns the states in the US with a population of at least 5 million can be expressed as:

BIN D ON COUNT(*) WHERE W = {State=AL,...,State=WY}
HAVING COUNT(*) > 5 million;

Note that since the answer to the query is a subset of the predicates in W (i.e. a subset of bin identifiers but not the aggregate values for these bins, an ICQ is not a linear counting query.

Top-k Counting Query (TCQ).

BIN D ON COUNT(*) WHERE $W = \{\phi_1, \dots, \phi_L\}$
ORDER BY COUNT(*) LIMIT k;

TCQ first sorts the bins based on their aggregate values (in descending order) and returns the top k bins identifiers (and not the aggregate values). For example, a query which returns the three US states with highest population can be expressed as:

BIN D ON COUNT(*) WHERE W = {State=AL,...,State=WY}
ORDER BY COUNT(*) LIMIT k;

3.2 Accuracy Measure

To ensure differential privacy, the answers to the exploration queries are typically noisy. To allow the data analyst to explore data with bounded error, we extend our queries to incorporate an accuracy requirement. The syntax for accuracy is inspired by that in BlinkDB [3]:

BIN D ON $f(\cdot)$ WHERE $W = \{\phi_1, \dots, \phi_L\}$
[HAVING $f(\cdot) > c$]
[ORDER BY $f(\cdot)$ LIMIT k]
ERROR α CONFIDENCE $1-\beta$;

We next define the semantics of the accuracy requirement for each of our query types. The accuracy requirement for a WCQ q_W is defined as a bound on the maximum error across queries in the workload W.

Definition 3.1 ((α,β)-WCQ accuracy). Given a workload counting query $q_W : \mathcal{D} \to \mathbb{R}^L$, where $W = \{\phi_1, \dots, \phi_L\}$. Let $M : \mathcal{D} \to \mathbb{R}^L$ be a mechanism that outputs a vector of answers y on D. Then, M satisfies (α,β)-W accuracy, if $\forall D \in \mathcal{D}$,

$$\Pr[\|y - q_W(D)\|_\infty \geq \alpha] \leq \beta, \qquad (2)$$

where $\|y - q_W(D)\|_\infty = \max_j |y[i] - c_{\phi_i}(D)|$.

The output of iceberg counting queries ICQ and top-k counting queries TCQ are not numeric, but a subset of the given workload predicates. Their accuracy measures are different from WCQ, and depend on their corresponding workload counting query q_W.

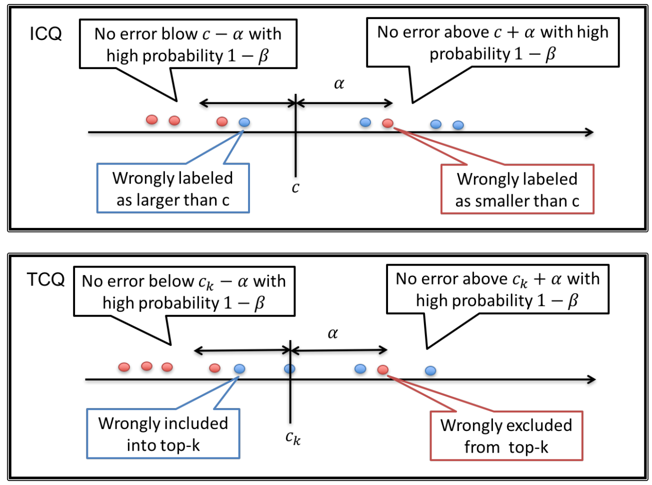

Figure 1: Accuracy Requirement for ICQ and TCQ

Definition 3.2 ((α,β)-ICQ accuracy). Given an iceberg counting query $q_{W,>c} : \mathcal{D} \to O$, where $W = \{\phi_1, \dots, \phi_L\}$, and O is a power set of W. Let $M : \mathcal{D} \to O$ be a mechanism that outputs a subset of W. Then, M satisfies (α,β)-ICQ accuracy for $q_{W,>c}$ if for D,

$$\Pr[|\{\phi \in M(D) \mid c_\phi(D) < c - \alpha\}| > 0] \leq \beta \qquad (3)$$

$$\Pr[|\{\phi \in (W - M(D)) \mid c_\phi(D) > c + \alpha\}| > 0] \leq \beta \qquad (4)$$

A mechanism for ICQ can make two kinds of errors: label predicates with true counts greater than c as $< c$ (red dots in Figure 1) and label predicates with true counts less than c as $> c$ (blue dots in Figure 1). We say a mechanism satisfies (α,β)-ICQ accuracy if with high probability, all the predicates with true counts greater than $c+\alpha$ are correctly labeled as $> c$ and all the predicates with true counts less than $c-\alpha$ are correctly labeled as $< c$. The mechanism may make arbitrary mistakes within the range $[c-\alpha, c+\alpha]$.

Definition 3.3 ((α,β)-TCQ accuracy). Given a top-k counting query $q_{W,k} : \mathcal{D} \to O$, where $W = \{\phi_1, \dots, \phi_L\}$, and O is a power set of W. Let $M : \mathcal{D} \to O$ be a mechanism that outputs a subset of W. Then, M satisfies (α,β)-TCQ accuracy if for $D \in \mathcal{D}$,

$$\Pr[|\{\phi \in M(D) \mid c_\phi(D) < c_k - \alpha\}| > 0] \leq \beta \qquad (5)$$

$$\Pr[|\{\phi \in (\Phi - M(D)) \mid c_\phi(D) > c_k + \alpha\}| > 0] \leq \beta \qquad (6)$$

where c_k is the k^{th} largest counting value among all the bins, and Φ is the true top-k bins.

The intuition behind Definition 3.3 is similar to that of ICQ and is explained in Figure 1: predicates with count greater than $c_k + \alpha$ are included and predicates with count less than $c_k - \alpha$ do not enter the top-k with high probability.

In the rest of the paper, we will describe how *APEx* designs differentially private mechanisms to answer queries with

Algorithm 1 *APEx Overview*

Require: Dataset D, privacy budget B

1: Initialize privacy loss $B_0 \leftarrow 0$, index $i \leftarrow 1$

2: **repeat**

3: Receive (q_i, α_i, β_i) from analyst

4: $\mathcal{M} \leftarrow$ mechanisms applicable to q_i's type

5: $\mathcal{M}^* \leftarrow \{M \in \mathcal{M} \mid M.\text{TRANSLATE}(q_i, \alpha_i, \beta_i).\epsilon^u \leq B - B_{i-1}\}$

6: **if** $\mathcal{M}^* \neq \emptyset$ **then**

7: *// Pessimistic Mode*

8: $M_i \leftarrow \texttt{argmin}_{M \in \mathcal{M}^*} M.\text{TRANSLATE}(q_i, \alpha_i, \beta_i).\epsilon^u$

9: *// Optimistic Mode*

10: $M_i \leftarrow \texttt{argmin}_{M \in \mathcal{M}^*} M.\text{TRANSLATE}(q_i, \alpha_i, \beta_i).\epsilon^l$

11: $(\omega_i, \epsilon_i) \leftarrow M_i.\text{RUN}(q_i, \alpha_i, \beta_i, D)$

12: $B_i \leftarrow B_{i-1} + \epsilon_i$, i++

13: **return** ω_i

14: **else**

15: $B_i = B_{i-1}$, i++

16: **return** 'Query Denied'

17: **end if**

18: **until** No more queries sent by local exploration

the above defined bounds on accuracy. The advantages of our accuracy definitions are that they are intuitive (when α increases, noisier answers are expected) and we can design privacy-preserving mechanisms that introduce noise while satisfying these accuracy guarantees. On the other hand, this measure is not equivalent to other bounds on accuracy like relative error and precision/recall which can be very sensitive to small amounts of noise (when the counts are small, or when lie within a small range). For instance, if the counts of all the predicates in ICQ lie outside $[c - \alpha, c + \alpha]$, a mechanism M that perturbs counts within $\pm\alpha$ and then answers an ICQ will have precision and recall of 1.0 with high probability as it makes no mistakes. However, if all the query answers lie within $[c - \alpha, c + \alpha]$, then the precision and recall of the output of M could be 0. Incorporating other error measures like precision/recall and relative error into *A*PEx is an interesting avenue for future work.

4 *A*PEX OVERVIEW

This section outlines how *A*PEx translates analyst queries with accuracy bounds into differentially private mechanisms, and how it ensures the privacy budget B specified by the data owner is not violated.

Accuracy Translator. Given an analyst's query (q, α, β), *A*PEx first uses the *accuracy translator* to choose a mechanism M that can (1) answer q under the specified accuracy bounds, with (2) minimal privacy loss. To achieve these, *A*PEx supports a set of differentially private mechanisms that can be used to answer each query type (WCQ, ICQ, TCQ). Multiple mechanisms are supported for each query type as different mechanisms result in the least privacy loss depending on the query and the dataset (as shown theoretically and empirically in Sections 5 and 7, respectively).

Each mechanism M exposes two functions: M.TRANSLATE, which *translates* a query and accuracy requirement into a lower and upper bound (ϵ^l, ϵ^u) on the privacy loss if M is executed, and M.RUN that runs the differentially private algorithm and returns an approximate answer ω for the query. The answer ω is guaranteed to satisfy the specified accuracy requirement. Moreover, M satisfies ϵ^u differential privacy. The mechanisms supported by *A*PEx and the corresponding TRANSLATE functions are described in Section 5. In some cases (e.g. Algorithm 4 in Section 5.3.2), the privacy loss incurred by M may be $\epsilon \in (\epsilon^l, \epsilon^u)$ that is smaller than the worst case, depending on the characteristics of the dataset.

As described in Algorithm 1, *A*PEx first identifies the mechanisms $\mathcal{M}$ that are applicable for the type of the query q_i (Line 4). Next, it runs M.TRANSLATE to get conservative estimates on privacy loss ϵ^u for all these mechanisms (Line 5). *A*PEx picks one of the mechanisms M from those that can be safely run using the remaining privacy budget, executes M.RUN, and returns the output to the analyst. As we will see there exist mechanisms where the privacy loss can vary based on the data in a range between $[\epsilon^l, \epsilon^u]$, and the actual privacy loss is unknown before running the mechanism. In such cases, *A*PEx can choose to be *pessimistic* and pick the mechanism with the least ϵ^u (Line 8), or choose to be *optimistic* and pick the mechanism with the least ϵ^l (Line 10).

Privacy Analyzer. Given a sequence of queries $(M_1, \ldots, M_i)$ already executed by the privacy engine that satisfy an overall B_{i-1}-differential privacy and a new query (q_i, α_i, β_i), *A*PEx identifies a set of mechanisms $\mathcal{M}^*$ that all will have a worst case privacy loss smaller than $B - B_{i-1}$ (Line 5). That is, running any mechanism in $\mathcal{M}^*$ will not result in exceeding the privacy budget in the worst case. If $\mathcal{M}^* = \emptyset$, then *A*PEx returns 'Query Denied' to the analyst (Line 16). Otherwise, *A*PEx runs one of the mechanisms M_i from $\mathcal{M}^*$ by executing M_i.RUN() and the output ω_i will be returned to the analyst. *A*PEx then increments B_{i-1} by the actual privacy loss ϵ_i rather than the upperbound ϵ^u (Line 12). As explained above, in some cases $\epsilon_i < \epsilon^u$ as different execution paths in the mechanism can have different privacy loss. Nevertheless, the privacy analyzer guarantees that the execution of any sequence of mechanisms $(M_1, M_2, \ldots, M_i)$ before it halts is B-differentially private (see Section 6, Theorem 6.2).

5 ACCURACY TRANSLATOR

In this section, we present the accuracy-to-privacy translation mechanisms supported by *A*PEx and the corresponding RUN and TRANSLATE functions. We first present a general baseline mechanism for all three types of exploration queries including workload counting query (WCQ), iceberg counting query (ICQ), and top-k counting query (TCQ). Then we show specialized mechanisms for each type of exploration queries which consumes smaller privacy cost under different scenarios than the baseline.

We represent the workload in these three queries in a matrix form, like the prior work for WCQ [26, 27, 31]. There are many possible ways to transform a workload into a matrix. Given a query with L predicates, the number of domain partitions can be as large as 2^L. In this work, we consider the following transformation to reduce complexity. Given a workload counting query q_W with the set of predicates $W = \{\phi_1, \ldots, \phi_L\}$, the full domain of the relation $dom(R)$ is partitioned based on W to form the new discretized domain $dom_W(R)$ such that any predicate $\phi_i \in W$ can be expressed as a union of partitions in the new domain $dom_W(R)$ and the number of partitions is minimized. For example, given $W = \{Age > 50 \wedge State = \text{AL}, \ldots, Age > 50 \wedge State = \text{WY}\}$, one possible partition is $dom_W(R) = \{Age > 50 \wedge State = \text{AL}, \ldots, Age > 50 \wedge State = \text{WY}, Age \leq 50\}$.

Let $\mathbf{x}$ represent the histogram of the table D over $dom_W(R)$. The set of corresponding counting queries $\{c_{\phi_1}, \ldots, c_{\phi_L}\}$ for q_W can be represented by a matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]^T$ of size $L \times |dom_W(R)|$. Hence, the answer to each counting query is $c_{\phi_i}(D) = \mathbf{w}_i \cdot \mathbf{x}$ and the answer to the workload counting query is simply $\mathbf{W}\mathbf{x}$. We denote and use this transformation by $\mathbf{W} \leftarrow \mathcal{T}(W), \mathbf{x} \leftarrow \mathcal{T}_W(D)$ throughout this paper. Unlike prior work [26, 27, 31] which aims to bound the expected total error for one query, *A*PEx aims to bound the maximum error per query with high probability which is more intuitive in the process of data exploration.

5.1 Baseline Translation

The baseline translation for all three query types is based on the Laplace mechanism [9, 10], a classic and widely used differentially private algorithm, which can be used for WCQ, ICQ, and TCQ. Formally,

Definition 5.1 (Laplace Mechanism (Vector Form)[9, 10]). Given an $L \times |dom_W(R)|$ query matrix $\mathbf{W}$, the randomized algorithm LM that outputs the following vector is ϵ-differentially private: $LM(\mathbf{W}, \mathbf{x}) = \mathbf{W}\mathbf{x} + Lap(b_{\mathbf{W}})^L$ where $b_{\mathbf{W}} = \frac{\|\mathbf{W}\|_1}{\epsilon}$, and $Lap(b)^L$ denote a vector of L independent samples η_i from a Laplace distribution with mean 0 and variance $2b^2$, i.e., $\Pr[\eta_i = z] \propto e^{-z/b}$ for $i = 1, \ldots, L$.

The constant $\|\mathbf{W}\|_1$ is equal to the sensitivity of queries set defined by the workload $\mathbf{W}$ [26, 27]. It measures the maximum difference in the answers to the queries in $\mathbf{W}$ on any two databases that differ only a single record. Mathematically, it is the maximum of $L1$ norm of a column of $\mathbf{W}$.

Algorithm 2 Laplace Mechanism (LM) (q, α, β, D)

1: Initialize $\mathbf{W} \leftarrow \mathcal{T}(W = \{\phi_1, \ldots, \phi_L\}), \mathbf{x} \leftarrow \mathcal{T}_W(D), \alpha, \beta$
2: **function** RUN(q, α, β, D)
3: $\quad \epsilon \leftarrow$ TRANSLATE$(q_W, \alpha, \beta).\epsilon^u$
4: $\quad [\tilde{x}_1, \ldots, \tilde{x}_L] \leftarrow \mathbf{W}\mathbf{x} + Lap(b)^L$, where $b = \|\mathbf{W}\|_1/\epsilon$
5: $\quad$ **if** q.type==WCQ (i.e., q_W) **then**
6: $\quad\quad$ **return** $([\tilde{x}_1, \ldots, \tilde{x}_L], \epsilon)$
7: $\quad$ **else if** q.type==ICQ (i.e., $q_{W,>c}$) **then**
8: $\quad\quad$ **return** $(\{\phi_i \in W \mid \tilde{x}_i > c\}, \epsilon)$
9: $\quad$ **else if** q.type==TCQ (i.e., $q_{W,k}$) **then**
10: $\quad\quad$ **return** $(\texttt{argmax}^k_{\phi_1, \ldots, \phi_L} \tilde{x}_i, \epsilon)$
11: $\quad$ **end if**
12: **end function**
13: **function** TRANSLATE(q, α, β)
14: $\quad$ **if** q.type==WCQ (i.e., q_W) **then**
15: $\quad\quad$ **return** $(\epsilon^u = \frac{\|\mathbf{W}\|_1 \ln(1/(1-(1-\beta)^{1/L}))}{\alpha}, \epsilon^l = \epsilon^u)$
16: $\quad$ **else if** q.type==ICQ (i.e., $q_{W,>c}$) **then**
17: $\quad\quad$ **return** $(\epsilon^u = \frac{\|\mathbf{W}\|_1 (\ln(1/(1-(1-\beta)^{1/L})) - \ln 2)}{\alpha}, \epsilon^l = \epsilon^u)$
18: $\quad$ **else if** q.type==TCQ (i.e., $q_{W,k}$) **then**
19: $\quad\quad$ **return** $(\epsilon^u = \frac{\|\mathbf{W}\|_1 2(\ln(L/(2\beta)))}{\alpha}, \epsilon^l = \epsilon^u)$
20: $\quad$ **end if**
21: **end function**

Algorithm 2 provides the RUN and TRANSLATE of Laplace mechanism for all three query types. This algorithm first transforms the query q_W and the data D into matrix representation $\mathbf{W}$ and $\mathbf{x}$. The TRANSLATE outputs a lower and upper bound (ϵ^l, ϵ^u) for each query type with a given accuracy requirement and these two bounds are the same as Laplace mechanism is data independent. However, these bounds vary among query types. The RUN takes the privacy budget computed by TRANSLATE(q, α, β) (Line 3) and adds the corresponding Laplace noise $[\tilde{x}_1, \ldots, \tilde{x}_L]$ to the true workload counts $\mathbf{W}\mathbf{x}$. When q is a WCQ, the noisy counts are returned directly; when q is an ICQ, the bin ids (the predicates) that have noisy counts $\geq c$ are returned; when q is a TCQ, the bin ids (the predicates) that have the largest k noisy counts are returned. Beside the noisy output, the privacy budget consumed by this mechanism is returned as well. The following theorem summarizes the properties of the two functions RUN and TRANSLATE.

THEOREM 5.2. *Given a query q where $q.type \in \{WCQ, ICQ, TCQ\}$, Laplace mechanism (Algorithm 2) denoted by M can achieve (α, β)-$q.type$ accuracy by executing the function* RUN(q, α, β, D) *for any $D \in \mathcal{D}$, and satisfy differential privacy with a minimal cost of* TRANSLATE$(q, \alpha, \beta).\epsilon^u$.

The accuracy and privacy proof is mainly based on the noise property of Laplace mechanism. Refer to Appendix A.1 for the detailed proof.

Algorithm 3 WCQ-SM $(q_W, \alpha, \beta, D, \mathbf{A})$

1: Initialize $\mathbf{W} \leftarrow \mathcal{T}(W), \mathbf{x} \leftarrow \mathcal{T}_W(D), \alpha, \beta, \mathbf{A}$
2: **function** RUN(q_W, α, β, D)
3: $\quad \epsilon \leftarrow$ TRANSLATE$(\mathbf{W}, \alpha, \beta).\epsilon^u$
4: $\quad \omega \leftarrow \mathbf{WA}^+(\mathbf{Ax} + Lap(b)^l)$, where $b = \|A\|_1/\epsilon$
5: $\quad$ **return** (ω, ϵ)
6: **end function**
7: **function** TRANSLATE(q_W, α, β)
8: $\quad$ Set $u = \frac{\|\mathbf{A}\|_1 \|\mathbf{WA}^+\|_F}{\alpha\sqrt{\beta/2}}$ and $l = 0$
9: $\quad \epsilon =$ BINARYSEARCH$(l, u,$ ESTIMATEBETA$(\cdot, \alpha, \beta, \mathbf{WA}^+))$
10: $\quad$ **return** $(\epsilon^u = \epsilon, \epsilon^l = \epsilon)$
11: **end function**
12: **function** ESTIMATEBETA$(\epsilon, \alpha, \beta, \mathbf{WA}^+)$
13: $\quad$ Sample size $N = 10000$ and failure counter $n_f = 0$
14: $\quad$ **for** $i \in [1, \ldots, N]$ **do**
15: $\quad\quad$ Sample noise $\boldsymbol{\eta}_i \sim Lap(\|\mathbf{A}\|_1/\epsilon)^l$
16: $\quad\quad$ **if** $\|(\mathbf{WA}^+)\boldsymbol{\eta}_i\|_\infty > \alpha$ **then**
17: $\quad\quad\quad n_f$++
18: $\quad\quad$ **end if**
19: $\quad$ **end for**
20: $\quad \beta_e = n_f/N, p = \beta/100$
21: $\quad \delta\beta = z_{1-p/2}\sqrt{\beta_e(1-\beta_e)/N}$
22: $\quad$ **return** $(\beta_e + \delta\beta + p/2) < \beta$
23: **end function**

5.2 Special Translation for WCQ

The privacy cost of the Laplace mechanism increases linearly with $\|\mathbf{W}\|_1$, the sensitivity of the workload in the query. For example, a prefix workload has a sensitivity equals to the workload size L. When L is very large, the privacy cost grows drastically. To address this problem, *A*PEx provides a special translation for WCQ, called *strategy-based mechanism*. This mechanism considers a different *strategy* workload $\mathbf{A}$ such that (i) $\mathbf{A}$ has a low sensitivity $\|\mathbf{A}\|_1$, and (ii) rows in $\mathbf{W}$ can be reconstructed using a small number of rows in $\mathbf{A}$.

Let strategy matrix $\mathbf{A}$ be a $l \times |dom_W(R)|$ matrix, and $\mathbf{A}^+$ denote its Moore-Penrose pseudoinverse, such that $\mathbf{WAA}^+ = \mathbf{W}$. Given such a strategy workload $\mathbf{A}$, we can first answer $\mathbf{A}$ using Laplace mechanism, i.e. $\hat{y} = \mathbf{Ax} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim Lap(b)^l$ and $b = \frac{\|\mathbf{A}\|_1}{\epsilon}$, and then reconstruct answers to $\mathbf{W}$ from the noisy answers to $\mathbf{A}$ (as a postprocessing step), i.e. $(\mathbf{WA}^+)\hat{y}$. This approach is formally known as the *Matrix mechanism* [26, 27] and shown in the RUN of Algorithm 3. If a strategy $\mathbf{A}$ is used for this mechanism, we denote it by $\mathbf{A}$-strategy mechanism. In this work, we consider several popular strategies in prior work [26, 27, 31], such as hierarchical matrix H_2. Techniques like HDMM [31] can automatically solve for a good strategy (but this is not our focus).

However, translating the accuracy requirement on WCQ-SM is nontrivial as the answers to the query $q_W(D)$ are linear combinations of noisy answers. The errors are due to the sums of weighted Laplace random variables which have non-trivial CDFs for the accuracy translation. Hence, we propose an accuracy to privacy translation method shown in the TRANSLATE function of Algorithm 3. We first set an upper bound u for the privacy to achieve (α, β)-WCQ accuracy based on Theorem A.1 (Appendix A.2) and conduct a binary search on a privacy cost ϵ between l and u (Line 9) such that the failing probability to bound the error by α equals to β. During this binary search, for each ϵ between l and u, we run Monte Carlo simulation to learn the empirical failing rate β_e to bound the error by α shown in the Function ESTIMATEBETA() such that with high confidence $1-p$, the true failing probability β_t lies within $\beta_e \pm \delta\beta$. If the empirical failing rate β_e is sufficiently smaller than β, then the upper bound is set to ϵ; otherwise the lower bound is set to ϵ. The next value for ϵ is $(l+u)/2$. This search process stops when l and u is sufficiently small. This approach can be generalized to all data-independent differentially private mechanisms. The simulation can be done offline.

THEOREM 5.3. *Given a workload counting query $q_W : \mathcal{D} \rightarrow \mathbb{R}^L$. Answering q_W with $\mathbf{A}$-strategy mechanism by executing RUN(q_W, α, β, D) in Algorithm 3 achieves (α, β)-WCQ accuracy for any $D \in \mathcal{D}$ with an approximated minimal privacy cost of TRANSLATE$(q_W, \alpha, \beta).\epsilon^u$.*

Refer to Appendix A.2 for the proof. In the future work, we can add similar optimizers like HDMM [31] which finds the optimal strategy efficiently for a given query.

5.3 Special Translation for ICQ

The strategy-based mechanism that we used for WCQ can be adapted to answer ICQ if used in conjunction with a post-processing step (Section 5.3.1). We also present another novel data dependent translation strategy for ICQ that may result in different privacy loss for different datasets given the same accuracy requirement (Section 5.3.2).

5.3.1 Strategy-based Mechanism (ICQ-SM). The analyst can pose a workload counting query q_W with (α, β)-WCQ requirement via *A*PEx, and then use the noisy answer of $q_W(D)$ to learn $q_{W,>c}(D)$ locally. This corresponds to a post-processing step of a differentially private mechanism and hence still ensures the same level of differential privacy guarantee (Theorem B.2). On the other hand, (α, β)-ICQ accuracy only requires to bound one-sided noise by α with probability $(1-\beta)$, and (α, β)-WCQ accuracy requires to bound two-sided noise with the same probability. Hence, if a mechanism has a failing probability of β to bound the error for WCQ, then using the same mechanism has a failing probability of $\beta/2$ to bound the error for ICQ.

5.3.2 Multi-Poking Mechanism (ICQ-MPM). We propose a data-dependent translation for ICQ, which can be used as a subroutine for mechanisms that involve threshold testing. For ease of explanation, we will illustrate this translation with a special case of ICQ when the workload size $L = 1$, denoted by, $q_{\phi,>c}(\cdot)$. Intuitively, when $c_\phi(D)$ is much larger (or smaller) than c, then a much larger (smaller resp.) noise can be added to $c_\phi(D)$ without changing the output of *A*PEx. Consider the following example.

Example 5.4. Consider a query $q_{\phi,>c}$, where $c = 100$. To achieve (α, β) accuracy for this query, where $\alpha = 10, \beta = 0.1^{10}$, the Laplace mechanism requires a privacy cost of $\frac{\ln(1/(2\beta))}{\alpha} = 2.23$ by Theorem 5.2, regardless of input D. Suppose $c_\phi(D) = 1000$. In this case, $c_\phi(D)$ is much larger than the threshold c, and the difference is $\frac{(1000-100)}{\alpha} = 90$ times of the accuracy bound $\alpha = 10$. Hence, even when applying Laplace comparison mechanism with a privacy cost equals to $\frac{2.23}{90} \approx 0.25$ wherein the noise added is bounded by 90α with high probability $1 - \beta$, the noisy difference $c_\phi(D) - c + \eta_{sign}$ will still be greater than 0 with high probability.

This is an example where a different mechanism rather than Laplace mechanism achieves the same accuracy with a smaller privacy cost. Note that the tightening of the privacy cost in this example requires to know the value of $c_\phi(D)$. It is difficult to determine a privacy budget for poking without looking at the query answer. To tackle this challenge, we propose an alternative approach that allows of m *pokes* with increasing privacy cost. This approach is summarized in Algorithm 4 as Multi-Poking Mechanism (MPM). This approach first computes the privacy cost if all m pokes are needed, $\epsilon_{\max} = \frac{\ln(m/(2\beta))}{\alpha}$. The first poke checks if bins have either sufficiently large noisy differences $\tilde{\mathbf{y}}$ with respect to the accuracy α_0 for the current privacy cost (Lines 8-10). If this is true (Lines 10), then the set of predicates with sufficiently large positive differences is returned; otherwise, the privacy budget is relaxed with additional $\epsilon_{\max}/m$. At $(i+1)$th iteration, instead of sampling independent noise, we apply the RelaxPrivacy Algorithm (details refer to [22]) to correlate the new noise $\boldsymbol{\eta}_{i+1}$ with noise $\boldsymbol{\eta}_i$ from the previous iteration. In this way, the privacy loss of the first $i + 1$ iterations is ϵ_{i+1}, and the noise added in the $i + 1$th iteration is equivalent to a noise generated with Laplace distribution with privacy parameter $b = (1/\epsilon_{i+1})$. This approach allows the data analyst to learn the query answer with a gradual relaxation of privacy cost. This process repeats until all $\epsilon_{\max}$ is spent. We show that Algorithm 4 achieves both accuracy and privacy requirements.

Theorem 5.5. *Given a query $q_{W,>c}$, Multi-Poking Mechanism (Algorithm 4), achieves (α, β)-ICQ accuracy by executing function* run$(q_{W,>c}, \alpha, \beta, D)$*, with differential privacy of* translate$(q_{W,>c}, \alpha, \beta).\epsilon^u$.

Algorithm 4 ICQ-MPM$(q_{W,>c}, \alpha, \beta, D, m)$

1: Initialize $\mathbf{W} \leftarrow \mathcal{T}(W), \mathbf{x} \leftarrow \mathcal{T}_W(D), \alpha, \beta, m = 10$
2: **function** run$(q_{W,>c}, \alpha, \beta, D)$
3: Compute $\epsilon_{\max}$ = translate$(q_{W,>c}, \alpha, \beta).\epsilon^u$
4: Initial privacy cost $\epsilon_0 = \epsilon_{\max}/m$
5: $\tilde{\mathbf{y}}_0 = \mathbf{W}\mathbf{x} - c + \boldsymbol{\eta}_0$, where $\boldsymbol{\eta}_0 \sim Lap(\|\mathbf{W}\|_1/\epsilon_0)^L$
6: **for** $i = 0, 1, \ldots, m-2$ **do**
7: Set $\alpha_i = \|\mathbf{W}\|_1 \ln(mL/(2\beta))/\epsilon_i$
8: $W_+ \leftarrow \{\phi_j \in W \mid (\tilde{\mathbf{y}}_i[j] - \alpha_i)/\alpha \geq -1\}$
9: $W_- \leftarrow \{\phi_j \in W \mid (\tilde{\mathbf{y}}_i[j] + \alpha_i)/\alpha \leq 1\}$
10: **if** $(W_+ \cup W_-) = W$ **then**
11: **return** (W_+, ϵ_i)
12: **else**
13: Increase privacy budget $\epsilon_{i+1} = \epsilon_i + \epsilon_{\max}/m$
14: **for** $j = 1, \ldots, L$ **do**
15: $\boldsymbol{\eta}_{i+1}[j]$ = RelaxPrivacy$(\boldsymbol{\eta}_i[j], \epsilon_i, \epsilon_{i+1})$ [22]
16: **end for**
17: New noisy difference $\tilde{\mathbf{y}}_{i+1} = \mathbf{W}\mathbf{x} - c + \boldsymbol{\eta}_{i+1}$
18: **end if**
19: **end for**
20: **return** $(\{\phi_j \in W \mid \tilde{\mathbf{y}}_{m-1}[j] > 0\}, \epsilon_{\max})$
21: **end function**
22: **function** translate$(q_{W,>c}, \alpha, \beta)$
23: **return** $\epsilon^u = \frac{\|\mathbf{W}\|_1 \ln(mL/(2\beta))}{\alpha}, \epsilon^l = \frac{\epsilon^u}{m}$
24: **end function**

Refer to Appendix A.3 for proof. The privacy loss of multi-poking mechanism at the worst case (the value returned by translate) is greater than that of the baseline LM, but this mechanism may stop before $\epsilon_{\max}$ is used up, and hence it potentially saves privacy budget for the subsequent queries.

5.4 Special Translation for TCQ

This section provides a translation mechanism, known as *Laplace top-k Mechanism* (shown in Algorithm 5). This mechanism is a generalized *report-noisy-max* algorithm [10]: when $k = 1$, it adds noise drawn from $Lap(1/\epsilon)$ to all queries, and only reports the query number that has the maximum noisy count (and not the noisy count). When $k \geq 1$, this mechanism first perturbs $\mathbf{W}\mathbf{x}$ with Laplace noise $\eta \sim Lap(b)^L$, where $b = k/\epsilon$. These predicates are then sorted based on their corresponding noisy counts in descending order, and the first k boolean formulae are outputted. The privacy cost is summarized in Theorem 5.6 and the proof follows that of the report-noisy-max algorithm as shown in Appendix A.4.

Theorem 5.6. *Given a top-k counting query $q_{W,k}(\cdot)$, where $W = \{\phi_1, \ldots, \phi_L\}$, for a table $D \in \mathcal{D}$, Laplace top-k mechanism (Algorithm 5) denoted by $LTM^{\alpha,\beta}_{W,k}(\cdot)$, can achieve (α, β)-TCQ accuracy by executing* run$(q_{W,k}, \alpha, \beta, D)$ *with minimal differential privacy cost of* translate$(q_{W,k}, \alpha, \beta).\epsilon^u$.

Note that the privacy proof of the report-noisy-max algorithm does not work for releasing both the noisy count and

Algorithm 5 TCQ-LTM($q_{W,k}, \alpha, \beta, D$))

Initialize $\mathbf{W} \leftarrow \mathcal{T}(W), \mathbf{x} \leftarrow \mathcal{T}_W(D), \alpha, \beta$
function RUN($q_{W,k}, \alpha, \beta, D$)
  $\epsilon \leftarrow$ TRANSLATE$(q_{W,k}, \alpha, \beta).\epsilon^u$
  $(\tilde{x}_1, \ldots, \tilde{x}_L) = \mathbf{W}\mathbf{x} + Lap(b)^L$, where $b = k/\epsilon$
  $(i_1, \ldots, i_k) = \mathtt{argmax}^k_{i=1,\ldots,L} \tilde{x}_i$
  return $(\{\phi_{i_1}, \ldots, \phi_{i_k}\}, \epsilon)$
end function
function TRANSLATE($q_{W,k}, \alpha, \beta$)
  return $\epsilon^u = \frac{2k \ln(L/(2\beta))}{\alpha}, \epsilon^l = \epsilon^u$
end function

the query number simultaneously. Hence we consider only releasing the bin identifiers in Algorithm 5. Moreover, the privacy cost of Algorithm 5 is independent of the workload $\|W\|_1$. On the other hand, the baseline LM (Algorithm 2) for answering TCQ queries is different from Algorithm 5. Algorithm 2 uses noise drawn from $Lap(\|W\|_1/\epsilon)$ to release noisy counts for all queries, and then picks the top-k as a post-processing step. Algorithm 2 allows the noisy counts to be shown to the analyst without hurting the privacy cost. Hence, Algorithm 2 has a simpler privacy proof than Algorithm 5 and a privacy loss that depends on the workload. *A*PEx supports both Algorithm 2 and Algorithm 5 as there is no clear winner between them when $k > 1$. *A*PEx chooses the one with the least epsilon for a given accuracy bound.

6 PRIVACY ANALYSIS

The privacy analyzer ensures that every sequence of queries answered by *A*PEx results in a B-differentially private execution, where B is the data owner specified privacy budget. The formal proof of privacy primarily follows from the well known composition theorems (described for completeness in Appendix B). According to sequential composition (Theorem B.1), the privacy loss of a set of differentially private mechanisms (that use independent random coins) is the sum of the privacy losses of each of these mechanisms. Moreover, postprocessing the outputs of a differentially private algorithm does not degrade privacy (Theorem B.2).

The main tricky (and novel) part of the privacy proof (described in Section 6.1) arises due to the fact that (1) the ϵ parameter for a mechanism is chosen based on the analyst's query and accuracy requirement, which in turn are adaptively chosen by the analyst based on previous queries and answers, and (2) some mechanisms may have an actual privacy loss that is dependent on the data. *A*PEx accounts for privacy based on the actual privacy loss (and not the worst case privacy loss (see Line 12, Algorithm 1).

6.1 Overall Privacy Guarantee

We show the following guarantee: any sequence of interactions between the data analyst and *A*PEx satisfies B-differential privacy, where B is the privacy budget specified by the data owner. In order to state this guarantee formally, we first need the notion of a *transcript of interaction* between *A*PEx and the data analyst.

We define the transcript of interaction $\mathbb{T}$ as an alternating sequence of queries (with accuracy requirements) posed to *A*PEx and answers returned by *A*PEx. $\mathbb{T}$ encodes the analyst's view of the private database. More formally,

- The transcript $\mathbb{T}_i$ after i interactions is a sequence $[(q_1, \alpha_1, \beta_1), (\omega_1, \epsilon_1), \ldots, (q_i, \alpha_i, \beta_i), (\omega_i, \epsilon_i)]$, where (q_i, α_i, β_i) are queries with accuracy requirements, and ω_i is the answer returned by *A*PEx and ϵ_i the actual privacy loss.
- Given $\mathbb{T}_{i-1}$, analyst chooses the next query $(q_{i+1}, \alpha_{i+1}, \beta_{i+1})$ adaptively. We model this using a (possibly randomized) algorithm $\mathbb{C}$ that maps a transcript $\mathbb{T}_{i-1}$ to (q_i, α_i, β_i); i.e., $\mathbb{C}(\mathbb{T}_{i-1}) = (q_i, \alpha_i, \beta_i)$. Note that the analyst's algorithm $\mathbb{C}$ does not access the private database D.
- Given (q_i, α_i, β_i), *A*PEx select a subset of mechanisms $\mathcal{M}^*$ such that $\forall M \in \mathcal{M}^*$, $M.$TRANSLATE$(q_i, \alpha_i, \beta_i).\epsilon^u \leq B - B_i$. Furthermore, if $\mathcal{M}^*$ is not empty, *A*PEx chooses one mechanism $M_i \in \mathcal{M}^*$ deterministically (either based on ϵ^l or ϵ^u) to run. The selection of M_i is deterministic and independent of D.
- If *A*PEx find no mechanism to run ($\mathcal{M}^* = \emptyset$), then the query is *declined* by *A*PEx. In this case, $\omega_i = \perp$ and $\epsilon_i = 0$.
- If the *A*PEx chosen algorithm M_i is LM, WCQ-SM, ICQ-SM or TCQ-LTM, $\epsilon_i = \epsilon_i^u$, where ϵ_i is the upperbound on the privacy loss returned by M_i.TRANSLATE. For ICQ-MPM, the actual privacy loss can be smaller; i.e., $\epsilon_i \leq \epsilon_i^u$.
- Let $Pr[\mathbb{T}_i|D]$ denote the probability that the transcript of interaction is $\mathbb{T}_i$ given input database D. The probability is over the randomness in the analyst's choices $\mathbb{C}$ and the randomness in the mechanisms $M_1, \ldots, M_i$ executed by *A*PEx.

Not all transcripts of interactions are realizable under *A*PEx. Given a privacy budget B, the set of valid transcripts is defined as:

Definition 6.1 (Valid Transcripts). A transcript of interaction $\mathbb{T}_i$ is a *valid* *A*PEx transcript generated by Algorithm 1 if given a privacy budget B the following conditions hold:

- $B_{i-1} = \sum_{j=1}^{i-1} \epsilon_j \leq B$, and
- Either $\omega_i = \perp$, or $B_{i-1} + \epsilon_i^u \leq B$.

We are now ready to state the privacy guarantee:

Theorem 6.2 (*APEx* Privacy Guarantee). *Given a privacy budget B, any valid APEx transcript* $\mathbb{T}_i$*, and any pair of databases* D, D' *that differ in one row (i.e.,* $|D \setminus D' \cup D' \setminus D| = 1$*), we have:*

(1) $B_i = \sum_{j=1}^{i} \epsilon_i \leq B$*, and*

(2) $Pr[\mathbb{T}_i|D] \leq e^{B_i} Pr[\mathbb{T}_i|D']$.

Proof. (1) directly follows from the definition of a valid transcript and these are the only kinds of transcripts an analyst sees when interacting with *A*PEx.

(2) can be shown as follows using induction.

Base Case: When the transcript is empty, $Pr[\emptyset|D] \leq e^0 Pr[\emptyset|D']$.

Induction step: Now suppose for all $\mathbb{T}_{i-1}$ of that encode valid *A*PEx transcripts of length $i-1$, $Pr[\mathbb{T}_{i-1}|D] \leq e^{B_{i-1}} Pr[\mathbb{T}_{i-1}|D']$. Let $\mathbb{T}_i = \mathbb{T}_{i-1}||[(q_i, \alpha_i, \beta_i), (\omega_i, \epsilon_i)]$ be a valid *A*PEx transcript of length i. Then:

$$\begin{aligned} Pr[\mathbb{T}_i|D] &= Pr[\mathbb{T}_{i-1}|D] Pr[[(q_i, \alpha_i, \beta_i), (\omega_i, \epsilon_i)]|D, \mathbb{T}_{i-1}] \\ &= Pr[\mathbb{T}_{i-1}|D] Pr[\mathbb{C}(\mathbb{T}_{i-1}) = (q_i, \alpha_i, \beta_i)] Pr[M_i(D) = (\omega_i, \epsilon_i)] \end{aligned}$$

Note that the analyst's choice of query q_i and accuracy requirement depends only on the transcript $\mathbb{T}_{i-1}$ and not the sensitive database, and thus incurs no privacy loss (from Theorem B.2). Thus, it is enough to show that

$$Pr[M_i(D) = (\omega_i, \epsilon_i)] \leq e^{\epsilon_i} Pr[M_i(D') = (\omega_i, \epsilon_i)]$$

Case 1: When $\omega_i \neq \perp$ and M_i is LM, WCQ-SM, ICQ-SM, or TCQ-LTM, the mechanism satisfies ϵ_i^u-DP and $\epsilon_i = \epsilon_i^u$. Therefore, $Pr[M_i(D) = (\omega_i, \epsilon_i)] \leq e^{\epsilon_i} Pr[M_i(D') = (\omega_i, \epsilon_i)]$.

Case 2: When $\omega_i \neq \perp$ and M_i is ICQ-MPM, the mechanism satisfies ϵ_i^u-DP across all outputs. However, when either mechanism outputs (ω_i, ϵ_i), for $\epsilon_i < \epsilon_i^u$, we can show that $Pr[M_i(D) = (\omega_i, \epsilon_i)] \leq e^{\epsilon_i} Pr[M_i(D') = (\omega_i, \epsilon_i)]$. In the case of ICQ-MPM, if the algorithm returns in Line 11 after i iterations of the loop, the noisy answer is generated by a DP algorithm with privacy loss $\epsilon_i = \frac{j}{m}\epsilon_i^u$.

Case 3: Finally, when $\omega_i = \perp$ (i.e., the query is declined), the decision to decline depends on ϵ_i^u of all mechanism applicable to the query (which is independent of the data) rather than ϵ_i (which could depend on the data in the case of ICQ-MPM). Therefore, $Pr[M_i(D) = (\omega_i, \epsilon_i)] = Pr[M_i(D') = (\omega_i, \epsilon_i)]$ for all D, D'. The proof would fail if the decision to deny a query depends on ϵ_i. □

7 QUERY BENCHMARK EVALUATION

In this section, we evaluate *A*PEx on real datasets using a set of benchmark queries. We show:

- *A*PEx is able to effectively translate queries associated with accuracy bounds into differentially private mechanisms. These mechanisms accurately answer a wide variety of interesting data exploration queries with moderate to low privacy loss.
- The set of query benchmarks show that no single mechanism can dominate the rest and *A*PEx picks the mechanism with the least privacy loss for all the queries.

7.1 Setup

Datasets. Our experiments use two real world datasets. The first data set `Adult` was extracted from 1994 US Census release [7]. This dataset includes 15 attributes (6 continuous and 9 categorical), such as "capital gain", "country", and a binary "label" indicating whether an individual earns more than 5000 or not, for a total of 32,561 individuals. The second dataset, refereed as `NYTaxi`, includes 9,710,124 NYC's yellow taxi trip records [1]. Each record consists of 17 attributes, such as categorical attributes (e.g., "pick-up-location"), and continuous attributes (e.g., "trip distance").

Query Benchmarks. We design 12 meaningful exploration queries on `Adult` and `NYTaxi` datasets, summarized in Table 1. These 12 queries cover the three types of exploration queries defined in Section 3.1, QW1-4, QI1-4, and QT1-4 corresponds to WCQ, ICQ, and TCQ respectively. Queries with number 1 and 2 are for `Adult`, and with number 3 and 4 are for `NYTaxi`. The predicate workload W cover 1D histogram, 1D prefix, 2D histogram and count over multiple dimensions. We set $\beta = 0.0005$ and vary $\alpha \in \{0.02, 0.04, 0.08, 0.16, 0.32, 0.64\}$.

Metrics. For each query (q, α, β), *A*PEx outputs (ϵ, ω) after running a differentially private mechanism, where ϵ is the actual privacy loss and ω is the noisy answer. The empirical error of a WCQ $q_W(D)$ is measured as $||\omega - q_W(D)||_\infty / |D|$, the scaled maximum error of the counts. The empirical errors of ICQ $q_{W,>c}(D)$ and TCQ $q_{W,k}(D)$ are measured as $||\alpha||_\infty/|D|$, the scaled maximum distance of mislabeled predicates.

Implementation Details. *A*PEx is implemented using python-3.4, and is run on a machine with 64 cores and 256 GB memory. We run *A*PEx with optimistic mode. For strategy mechanism, we choose H_2 strategy (a hierarchical set of counts [26, 27, 31]) for all queries.

7.2 *A*PEx End-to-End Study

We run *A*PEx for the 12 queries shown in Table 1 with different accuracy requirements from $0.01|D|$ to $0.64|D|$ and $\beta = 0.0005$. We show in Figure 2 that a line connects points (α, ϵ^u) where ϵ^u is the upper bound on the privacy loss for the mechanism chosen by *A*PEx for the given α. For all the queries except QI2 and QI3, the mechanism chosen for each α incurs an actual privacy cost at $\epsilon = \epsilon^u$ and the only variation in the empirical error, so the corresponding $(\hat{\alpha}/|D|, \epsilon)$ of 10 runs is shown as boxplots. For QI2 and QI3, both the empirical error and the actual privacy cost $(\hat{\alpha}/|D|, \epsilon)$ vary across runs and hence are plotted as points in Figure 2.

Name	D	Query workload W	Query output
QW1	Adult	"capital gain"$\in [0, 50)$, "capital gain"$\in [50, 100)$, ...,"capital gain"$\in [4950, 5000)$	bin counts
QW2	Adult	"capital gain"$\in [0, 50)$, "capital gain"$\in [0, 100)$, ..., "capital gain"$\in [0, 5000)$	bin counts
QW3	NYTaxi	"trip distance"$\in [0, 0.1)$, "capital gain"$\in [0, 50)$, ...,"capital gain"$\in [0, 50)$	bin counts
QW4	NYTaxi	$(0 \leq$"total amount"$< 1\wedge$ "passenger"$= 1)$,..., $(9 \leq$"total amount"$< 10\wedge$ "passenger"$= 10)$	bin counts
QI1	Adult	"capital gain"< 50, "capital gain"< 100,..., "capital gain"< 5000	bin ids having counts $> 0.1\|D\|$
QI2	Adult	$(0 \leq$"capital gain"< 100, "sex"='M'),...(4500 $\leq$"capital gain"< 5000, "sex"='F')	bin ids having counts $> 0.1\|D\|$
QI3	NYTaxi	"fare amount"$\in [0, 0, 1)$, "fare amount"$\in [0.1, 2)$,..., "fare amount"$\in [9.9, 10)$	bin ids having counts $> 0.1\|D\|$
QI4	NYTaxi	"total amount"$\in [0, 0, 1)$, "total amount"$\in [0.1, 2)$,..., "total amount"$\in [9.9, 10)$	bin ids having counts $> 0.1\|D\|$
QT1	Adult	"age"$= 0$,"age"$= 1$,...,"age"$= 99$	top 10 bins with highest counts
QT2	Adult	100 predicates on different attributes, e.g. "age"$= 1$, "workclass"="private",...	top 10 bins with highest counts
QT3	NYTaxi	("PUID"=1 $\wedge$"DOID"=1), ("PUID"=1 $\wedge$ "DOID"=2),...,("PUID"=10 $\wedge$ "DOID"=10)	top 10 bins with highest counts
QT4	NYTaxi	100 predicates on different attributes, e.g. "pickup date"$= 1$, "passenger count"$= 1$,...	top 10 bins with highest counts

Table 1: Query benchmarks includes 3 types of exploration queries on 2 datasets.

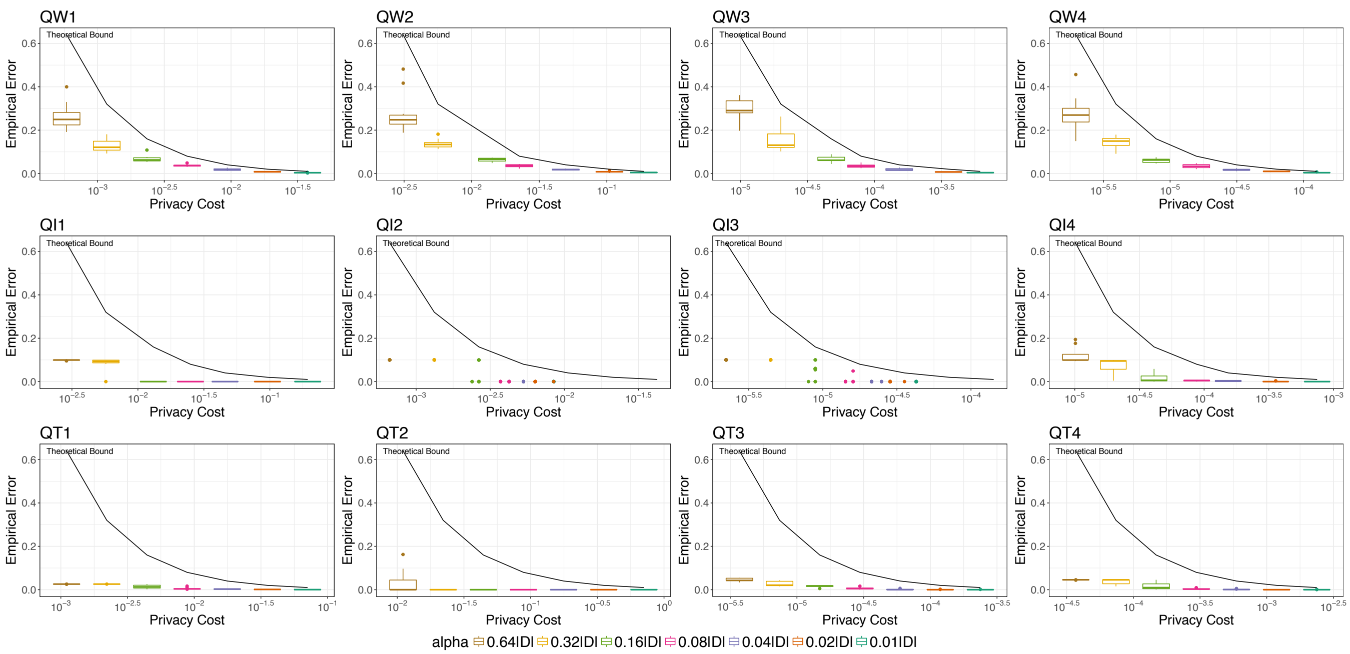

Figure 2: Privacy cost and empirical accuracy using optimal mechanism chosen by APEx (Optimistic Mode) on the 12 queries at default parameter setting with $\alpha \in \{0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64\}|D|$ and $\beta = 5 \times 10^{-4}$. On Adult data, all queries can be answered with empirical error < 0.1 with privacy budget < 0.1; on NYTaxi data, all queries can be answered with empirical error < 0.1 with privacy budget < 0.001. When accuracy requirement relaxes (i.e., α increases), the privacy cost decreaess and the empirical accuracy decreases for all queries.

The empirical error α is always bounded by the theoretical α for all the queries. The gap between the theoretical line and the actual boxplots/points are: (1) the analysis of the error is not tight due to the use of union bound; (2) for mechanism with data dependent translation (QI2 and QI3), the actual privacy cost is far from the upperbound ϵ^u resulting a left shift of the points from the theoretical line. The privacy cost for QW1 and QW2 for Adult dataset is in the range of $(10^{-4}, 10^{-1})$ for all α values. This privacy cost is 2-3 orders larger than the privacy cost for QW3 and QW4 on the NYTaxi dataset, because given the same ratio $\alpha/|D|$, the queries on NYTaxi has a larger α than Adult because of data size.

Additional experiments using F1 score as the quality measurement are included in Appendix C.

Mechanism	Query-α							
	QW1-0.02|D|	QW1-0.08|D|	QW2-0.02|D|	QW2-0.08|D|	QW3-0.02|D|	QW3-0.08|D|	QW4-0.02|D|	QW4-0.08|D|
WCQ-LM	**0.01874**	**0.00469**	1.87430	0.46858	0.00629	0.00157	**0.00006**	**0.00002**
WCQ-SM	0.09880	0.02383	**0.10451**	**0.02251**	**0.00036**	**0.00008**	0.00033	0.00009
	QI1-0.02|D|	QI1-0.08|D|	QI2-0.02|D|	QI2-0.08|D|	QI3-0.02|D|	QI3-0.08|D|	QI4-0.02|D|	QI4-0.08|D|
ICQ-LM	1.76786	0.44197	0.01768	0.00442	0.00006	0.000015	0.00593	0.00148
ICQ-SM	**0.10271**	**0.02682**	0.10506	0.02517	0.00033	0.00008	**0.00034**	**0.00008**
ICQ-MPM	0.63644	0.31822	**0.00636**	**0.00371**	**0.00003**	**0.000014**	0.00640	0.00178
	QT1-0.02|D|	QT1-0.08|D|	QT2-0.02|D|	QT2-0.08|D|	QT3-0.02|D|	QT3-0.08|D|	QT4-0.02|D|	QT4-0.08|D|
TCQ-LM	**0.03536**	**0.00884**	266.24590	66.56148	**0.00012**	**0.00003**	1.40857	0.35214
TCQ-LTM	0.35358	0.08840	**0.35358**	**0.08840**	0.00119	0.00030	**0.00119**	**0.00030**

Table 2: Privacy cost using all applicable mechanisms on the 12 queries (Table 1) at $\alpha = \{0.02, 0.08\}|D|$ and $\beta = 5 \times 10^{-4}$. Median of 10 runs is reported for data-dependent mechanisms. It shows that (a) no single mechanism can always win or lose, (b) privacy cost of different mechanisms answering the same query, and privacy cost of same mechanism on different queries can be significantly different. Therefore, it is critical to use *A*PEx for choosing optimal mechanisms.

7.3 Optimal Mechanism Study

We run all the applicable mechanisms for the 12 queries from Table 1 at $\alpha \in \{0.02|D|, 0.08|D|\}$ and show the median of the actual privacy costs in Table 2. The privacy cost with the least value is in bold for the given query and accuracy. Indeed, *A*PEx picks the mechanisms with these least privacy cost for all the 12 queries. *A*PEx can save more than 90% of the privacy cost of the baseline translation (LM), such as QW2, QW3, QT2,QT4, and all the ICQ. In particular, the baseline mechanism LM is highly dependent on the sensitivity of the query. For example, the workload in QW2 (a cumulative histogram query) and QT2 (counts on many attributes) has a high sensitivity, so the cost of WCQ-LM is 20 times larger than WCQ-SM for QW2 at $\alpha = 0.08|D|$, and the cost of TCQ-LM is 760X more expensive than TCQ-LTM for QT2 at $\alpha = 0.02|D$, where WCQ-SM and TCQ-LTM are the optimal mechanisms chosen by *A*PEx for QW2 and QT2 respectively.

The optimal mechanism chosen by *A*PEx can change query parameters: workload size L, threshold c in ICQ, and k in TCQ. As parameter L and k play a direct role in the privacy cost shown in Section 5, we will focus on the sensitivity of the privacy on c and leave the rest in the full version [12].

Figure 3 shows the actual privacy cost of mechanisms for QI2 with different thresholds. All the mechanisms for ICQ except ICQ-MPM, have a fixed privacy cost which is dependent of the data and the query (including c). However, we observe an interesting trend for the actual privacy cost used by ICQ-MPM as c increases. The smallest privacy cost which takes place after $c = 0.8$ is $1/10$ of the upper bound of ICQ-MPM. The privacy cost of ICQ-MPM depends on the number of poking times before returning a output, which is related to the distance between the threshold and the true counts associated to the predicates. If the predicates are far from the threshold, the fewer number of poking is required and hence a smaller privacy budget is spent. On the other hand, if the true count is very close to the threshold, then it requires a small noise and hence all the budget to decide the label of this predicate with confidence. When $c = 0.01|D|$, 98% predicates are within the range $[c - \alpha, c + \alpha]$, and hence to confidently decide the label for all these predicates require more poking and hence a larger privacy cost. Consider the many predicates having counts close to $0.01|D|$, the cost of ICQ-MPM is high. As c increases to $0.10|D|$, all predicates have sufficient different counts as c, then 1 or 2 times of poking are sufficient. When c continue increases to $0.32|D|$, there is a single predicate that with true count (which is $0.3117|D|$) closer to c , it again requires more pokings to make a confident decision. A similar behavior is seen when c is around $0.6050|D|$.

Moreover, in the bad cases where c is close to true counts, the actual privacy cost of ICQ-MPM might be more expensive

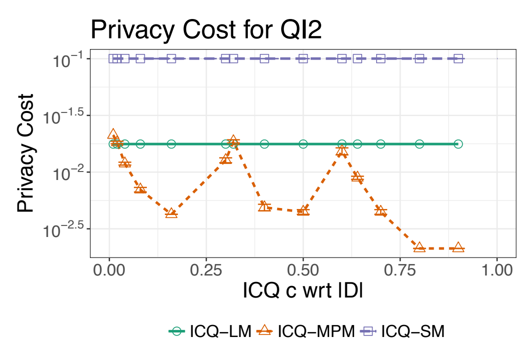

Figure 3: Vary ICQ c

than the baseline ICQ-LM. For example, when $c = 0.01|D|$, ICQ-LM is better. This is a case where *A*PEx under optimistic mode fails to choose ICQ-LM as the optimal mechanism.

In summary, we see that the optimal mechanism with the least privacy cost changes among queries and even the same query with different parameters. This shows there is a great need for systems like *A*PEx to provide translation and identify optimal mechanisms.

8 CASE STUDY

In this section, we design an application benchmark based on entity resolution to show that (1) we can express real data exploration workflow using our exploration language; (2) *A*PEx allows entity resolution workflow to be conducted with high accuracy and strong privacy guarantee, (3) *A*PEx allows trade-off between privacy and final task quality for a given exploration task.

8.1 Case Study Setting

Entity Resolution (ER) is an application of identifying table records that refer to the same real-world object. In this case, we use the `citations` [6] dataset to perform exploration tasks. Each row in the table is a pair of citation records, with a binary label indicating whether they are duplicates or not. All the citation records share the same schema, which consists of 3 text attributes of title, authors and venue, and one integer attribute of publication year. A training set D of size 4000 is sampled from `citations` such that every record appears at most once. Two exploration tasks for entity resolution on D are considered: *blocking* and *matching*. These two tasks are achieved by learning a boolean formula P (e.g. in DNF) over *similarity predicates*. We express a similarity predicate p as a tuple $(A, t, sim, \theta) \in attr(R) \times \mathbb{T} \times \mathbb{S} \times \Theta$, where $A \in attr(R)$ is an attribute in the schema $t \in \mathbb{T}$ is a transformation of the attribute value, and $sim \in \mathbb{S}$ is a similarity function that usually takes a real value often restricted to $[0, 1]$, and $\theta \in \Theta$ is a threshold. Given a pair of records (r_1, r_2), a similarity predicate p returns either 'True' or 'False' with semantics: $p(r_1, r_2) \equiv (sim(t(r_1.A), t(r_2.A)) > \theta)$.

In this case study, the exploration task for blocking is to find a boolean formula P_b that identifies a small set of candidate matching pairs that cover most of the true matches, known as high *recall*, with a small blocking cost (a small fraction of pairs in the data that return true for P_b). The exploration task for matching is to identify a boolean formula P_m that identifies matching records that achieves high *recall* and *precision* on D_t. Precision measures whether P_m classifies true non-matches as matches, and recall measures the fraction of true matches that are captured by P_m. The quality of this task is measured by $F_1^{P_m}$, the harmonic mean of precision and recall.

Using exploration queries supported by *A*PEx, we can express two exploration strategies for each task: BS1 (MS1) using WCQ only to complete blocking (matching) task, and BS2 (MS2) using ICQ/TCQ to conduct the blocking (matching) task. Each strategy generates a sequence of exploration queries which interact with *A*PEx and constructs a boolean formula for the corresponding task. For each strategy, we randomly sample a concrete cleaner from our cleaner model and report mean and quartiles of its cleaning quality over 100 runs. Both the cleaner model and the exploration strategies are detailed in the full version [12].

8.2 End-to-End Task Evaluation

We report the end-to-end performance of *A*PEx on the four exploration strategies.

Vary Privacy Constraint. Given an exploration strategy from BS1, BS2, MS1, MS2 on training data D, a sequence of queries is generated based on its interaction with *A*PEx until the privacy loss exceeds the privacy constraint B specified by the data owner. The accuracy requirement for this set of experiments is fixed to be $\alpha = 0.008|D_t|, \beta = 0.0005$. The privacy constraint varies from 0.1 to 2. Each exploration strategy is repeated 100 times under a given privacy constraint and we report the output quality of these 100 runs at different privacy constraint.

Figure 4 shows the exploration quality (recall for blocking task and F1 for matching task) of 100 runs of each exploration strategy at $B = \{0.1, 0.2, 0.5, 1, 1.5, 2\}$. We observe that the expected task quality (median and variance) improves as the budget constraint increases and gets stable after, for example reaching $B \geq 1$ for MS1. Since we fixed α, the privacy loss for each exploration query is also fixed. Thus, B directly controls the number of queries *A*PEx answers before halting. For small $B < 0.2$, only a few queries are answered and the cleaning quality is close to random guessing. As B increases, more queries can be learned about the dataset. After B reaches a certain value around 1.0, *A*PEx can answer sufficient number of queries; therefore obtaining high accuracy. For MS2, when it reaches good F1 score at $B = 1$, more noisy answers can mislead the MS2 strategy to add more predicates to the blocking conjunction function, and decrease the quality.

The privacy cost used by each ICQ and TCQ is generally less than what has spent on the corresponding WCQ, as less information is shown to the data analyst. Given the same privacy budget, e.g. when $B = 0.5$, more queries can be answered in BS2 than BS1, results in a 25% better recall. We observe the same trend in MS1 and MS2. This shows that it is important for *A*PEx to support translation for different types of queries for better end-to-end accuracy.

Vary Accuracy Requirement. This section shows the task quality of each exploration strategy at different accuracy

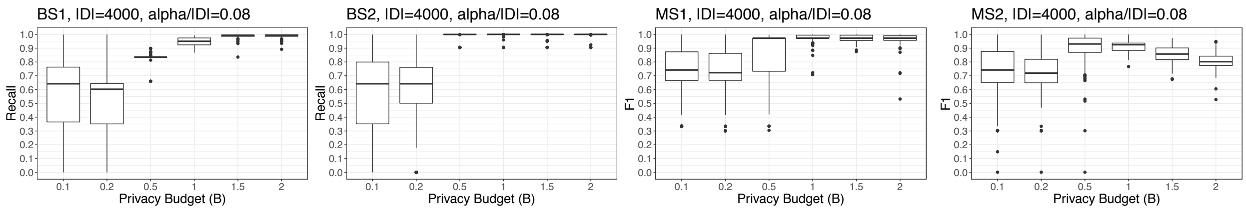

Figure 4: Performance of *A*PEx for blocking (BS1, BS2) and matching (MS1, MS2) tasks with increasing privacy budget B at fixed $\alpha = 0.08|D|$: the expected task quality improves with smaller variance as the budget constraint increases and gets stable. Fixing α fixes privacy cost per operation. Thus increasing B increases the number of queries answered.

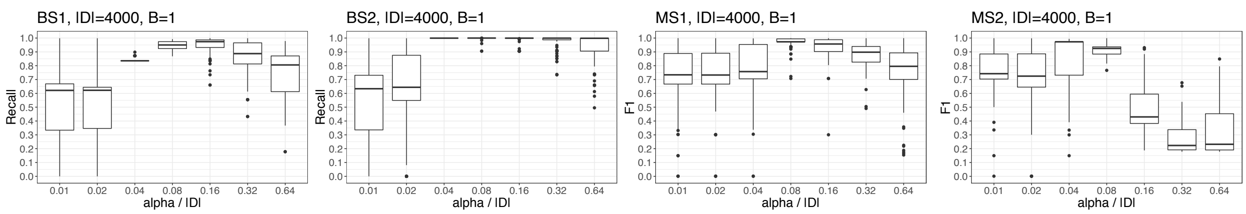

Figure 5: Performance of *A*PEx for blocking (BS1, BS2) and matching (MS1, MS2) at fixed privacy budget $B = 1$ with increasing α from $0.01|D_t|$ to $0.64|D_t|$. There exists an optimal α to achieve highest quality at a given privacy constraint. Increasing α decreases privacy cost per operation. Thus for a fixed budget this increases the number of queries. However, many queries each with a low privacy budget is not good for end-to-end accuracy.

requirements under a fixed privacy constraint $B = 1.0$. For each $\alpha \in \{0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64\}|D|$, the exploration strategy interacts with *A*PEx until the privacy loss exceeds $B = 1.0$. Each exploration is repeated 100 times and we report the quality of the constructed boolean formula of these 100 runs. As shown in Figure 5 the quality of the four exploration strategies all improves first as accuracy requirement relaxes and then degrades again under $B = 1.0$. This is because when there is a privacy constraint, *A*PEx only allows a limited number of queries to be answered. Given this fixed privacy budget $B = 1$ and a fixed accuracy requirement α, the number of queries can be answered is bounded and related to α. When α first increases from $0.01|D|$ to $0.08|D|$, more queries can be answered and give more information on which predicates to choose. However, as α keeps growing, the answers get noisier and misleading, resulting in the drop in quality, even though more questions can be answered. Hence, choosing the optimal α for different queries in an exploration process is an interesting direction for future work.

9 CONCLUSIONS

We proposed *A*PEx, a framework that allows data analysts to interact with sensitive data while ensuring that their interactions satisfy differential privacy. Using experiments with query benchmarks and entity resolution application, we established that *A*PEx allows high exploration quality with a reasonable privacy loss.

*A*PEx opens many interesting future research directions. First, more functionalities can be added to *A*PEx: (a) a recommender which predicts the subsequent interesting queries and advises the privacy cost of these queries to the analyst; (b) an inferencer which uses historical answers to reduce the privacy cost of a new incoming query; and (c) a sampler which incorporate approximate query processing to have 3-way trade-off between efficiency, accuracy, and privacy. More translation algorithms from accuracy to privacy, especially for data-dependent mechanisms and non-linear queries can be implemented in *A*PEx to further save privacy budget throughout the exploration. We show the exploration queries in *A*PEx to support entity resolution tasks. These exploration queries can be extended to support other exploration tasks, e.g., schema matching, feature engineering, tuning machine learning workloads. This would also require extending our system to handle relations with multiple tables, constraints like functional dependencies, etc. *A*PEx turns the differentially private algorithm design problem on its head – it minimizes privacy loss given an accuracy constraint. This new formulation has applications on sensitive data exploration and can trigger an entirely new line of research.

Acknowledgments: This work was supported by the National Science Foundation under grants 1253327, 1408982; by DARPA and SPAWAR under contract N66001-15-C-4067; by NSERC Discovery grant; and by Thomson Reuters-NSERC IRC grant.

REFERENCES

[1] Tlc trip record data. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

[2] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *OJ*, 2016-04-27.

[3] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. Blinkdb: Queries with bounded errors and bounded response times on very large data. In *EuroSys*, 2013.

[4] C. Binnig, J. Freire, and E. Wu, editors. *HILDA@SIGMOD*. ACM, 2018.

[5] F. Chiang and D. Gairola. Infoclean: Protecting sensitive information in data cleaning. *J. Data and Information Quality*, 2018.

[6] S. Das, A. Doan, P. S. G. C., C. Gokhale, and P. Konda. The magellan data repository. https://sites.google.com/site/anhaidgroup/projects/data.

[7] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.

[8] C. Dwork. Differential privacy. In *IN ICALP*. Springer, 2006.

[9] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[10] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 2014.

[11] Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *CCS*, 2014.

[12] C. Ge, I. Ilyas, X. He, and A. Machanavajjhala. Private exploration primitives for data cleaning. *CoRR*, abs/1712.10266, 2017.

[13] A. Greenberg. Apple's 'differential privacy' is about collecting your data—but not *your* data. *Wired*, 2016.

[14] S. Haney, A. Machanavajjhala, J. Abowd, M. Graham, M. Kutzbach, and L. Vilhuber. Utility cost of formal privacy for releasing national employer-employee statistics. In *SIGMOD*, 2017.

[15] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In *NIPS*, 2012.

[16] M. Hay, A. Machanavajjhala, G. Miklau, Y. Chen, and D. Zhang. Principled evaluation of differentially private algorithms using dpbench. In *SIGMOD*, 2016.

[17] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *KDD*, 2013.

[18] N. M. Johnson, J. P. Near, and D. X. Song. Practical differential privacy for SQL queries using elastic sensitivity. *CoRR*, abs/1706.09479, 2017.

[19] G. King. Social science one: Public launch, July 2018.

[20] G. King and N. Persily. A new model for industry-academic partnerships, 2018 Working Paper.

[21] P. Konda, S. Das, P. S. G. C., A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. F. Naughton, S. Prasad, G. Krishnan, R. Deep, and V. Raghavendra. Magellan: Toward building entity matching management systems. *PVLDB*, 2016.

[22] F. Koufogiannis, S. Han, and G. J. Pappas. Gradual release of sensitive data under differential privacy. *CoRR*, abs/1504.00429, 2015.

[23] S. Krishnan, J. Wang, M. J. Franklin, K. Goldberg, and T. Kraska. Privateclean: Data cleaning and differential privacy. In *SIGMOD*, 2016.

[24] J. Lee and C. W. Clifton. Top-k frequent itemsets via differentially private fp-trees. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[25] C. Li, M. Hay, G. Miklau, and Y. Wang. A data- and workload-aware algorithm for range queries under differential privacy. *Proc. VLDB Endow.*, 2014.

[26] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *PODS*, 2010.

[27] C. Li, G. Miklau, M. Hay, A. Mcgregor, and V. Rastogi. The matrix mechanism: Optimizing linear counting queries under differential privacy. *The VLDB Journal*, 2015.

[28] N. Li, W. Qardaji, D. Su, and J. Cao. Privbasis: Frequent itemset mining with differential privacy. *Proc. VLDB Endow.*, 2012.

[29] K. Ligett, S. Neel, A. Roth, B. Waggoner, and Z. S. Wu. Accuracy first: Selecting a differential privacy level for accuracy-constrained erm. In *NIPS*, 2017.

[30] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *IEEE ICDE*, 2008.

[31] R. McKenna, G. Miklau, M. Hay, and A. Machanavajjhala. Optimizing error of high-dimensional statistical queries under differential privacy. In *VLDB*, 2018.

[32] F. D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *SIGMOD*, 2009.

[33] D. Proserpio, S. Goldberg, and F. McSherry. Calibrating data to sensitivity in private data analysis: A platform for differentially-private analysis of weighted datasets. *Proc. VLDB Endow.*, 2014.

[34] M. Stonebraker, D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. B. Zdonik, A. Pagan, and S. Xu. Data curation at scale: The data tamer system. In *CIDR*, 2013.

[35] L. Vilhuber, I. M. Schmutte, and J. M. Abowd. Proceedings from the 2016 NSF–Sloan workshop on practical privacy, Jan 2017.

[36] C. Zeng, J. F. Naughton, and J.-Y. Cai. On differentially private frequent itemset mining. *VLDB*, 2012.

[37] D. Zhang, R. McKenna, I. Kotsogiannis, M. Hay, A. Machanavajjhala, and G. Miklau. Ektelo: A framework for defining differentially-private computations. In *SIGMOD*, 2018.

A THEOREMS AND PROOFS

A.1 Laplace Mechanism (Theorem 5.2)

Privacy Proof for WCQ

Proof: For any pair of neighbouring databases D and D' such that $|D \backslash D' \cup D' \backslash D| = 1$, given a WCQ query q_W, LM adds noise drawn from $Lap(b)$ to query c_{ϕ_i}, where $b = \|\mathbf{W}\|_1/\epsilon$. Consider a column vector of counts $y = [y_1, \cdots, y_L]$,

$$\frac{\Pr[q_{\mathbf{W}}(D) + Lap(b_{\mathbf{W}})^L = y]}{\Pr[q_{\mathbf{W}}(D') + Lap(b_{\mathbf{W}})^L = y]} \leq exp(\epsilon)$$

Therefore, it satisfies $\epsilon = \|\mathbf{W}\|_1/b$-differential privacy. □

Privacy Proof for ICQ and TCQ

Proof: For ICQ (and TCQ, respectively), Line 8 (Line 10) post-processes the noisy answers only without accessing the true data. By the post-processing property of differential privacy (Theorem B.2), Laplace mechanism for ICQ (TCQ) satisfies $\epsilon = \|\mathbf{W}\|_1/b$-differential privacy. □

Accuracy Proof for WCQ

Proof: Give a WCQ q_W, for any $D \in \mathcal{D}$, setting $b \leq \frac{\alpha}{\ln(1/(1-(1-\beta)^{1/L}))}$ bounds the failing probability, i.e.,

$$\begin{aligned}&\Pr[\|LM(\mathbf{W}, \mathbf{x}) - q_W(D)\|_\infty \geq \alpha] \\ &= \quad 1 - \prod_{i \in [1,L]} (1 - \Pr[|\eta_i| > \alpha]) = 1 - (1 - e^{-\alpha/b})^L \leq \beta \quad \square\end{aligned}$$

Accuracy Proof for ICQ

Proof: Given ICQ $q_{W,>c}$, for any $D \in \mathcal{D}$, setting $b \leq$

$\frac{\alpha}{\ln(1/(1-(1-\beta)^{1/L}))-\ln 2}$ bounds the failure probability, i.e.,

$$\begin{aligned} & \Pr[|\{\phi \in M(D) \mid c_\phi(D) < c - \alpha\}| > 0] \\ = \;& 1 - \prod_{\phi \in W: c_\phi(D) - c + \alpha < 0} (1 - \Pr[c_\phi(D) - c + \eta > 0]) \\ \leq \;& 1 - \prod_{\phi \in W: c_\phi(D) - c + \alpha < 0} (1 - \Pr[\eta > \alpha]) \\ \leq \;& 1 - (1 - e^{-\alpha/b}/2)^L < \beta \end{aligned}$$

The proof for the other condition is analogous. □

Accuracy Proof for TCQ

Proof: Given a TCQ $q_{W,k}$, for any $D \in \mathcal{D}$, W.L.O.G. let $c_{\phi_1}(D) \geq \cdots \geq c_{\phi_k}(D) \cdots \geq c_L(D)$ and the noise added to these counts be $\eta_1, \ldots, \eta_L$ respectively. Let $c_k = c_{\phi_k}(D)$ the k^{th} largest counting value. Setting noise parameter $b \leq \frac{\alpha}{2\ln(L/(2\beta))}$ bounds the failing probability, i.e.,

$$\begin{aligned} & \Pr[|\{\phi \in M(D) \mid c_\phi(D) < c_k - \alpha\}| > 0] \\ \leq \;& 1 - \Pr[(\max_{i > k, c_{\phi_i}(D) < c_k - \alpha} \eta_i < \frac{\alpha}{2}) \wedge (\min_{i \leq k} \eta_i > -\frac{\alpha}{2})] \\ = \;& \Pr[(\max_{i > k, c_{\phi_i}(D) < c_k - \alpha} \eta_i \geq \frac{\alpha}{2}) \vee (\min_{i \leq k} \eta_i \geq -\frac{\alpha}{2})] \\ \leq \;& (L-k)e^{-\alpha/(2b)}/2 + ke^{-\alpha/(2b)}/2 = Le^{\alpha/(2b)}/2 \leq \beta \quad (7) \end{aligned}$$

The proof for $\Pr[|\{\phi \in (\Phi - M(D)) \mid c_\phi(D) > c_k + \alpha\}| > 0] \leq \beta$ is analogous. □

A.2 Strategy-based Mechanism for WCQ

Theorem A.1. *Given a WCQ $q_W : \mathcal{D} \to \mathbb{R}^L$, for a table $D \in \mathcal{D}$, let $(\mathbf{W}, \mathbf{x}) = \mathcal{T}(W, D)$. Let $\mathbf{A}$ be a strategy used in Algorithm 3 to answer q_W. When $\epsilon \geq \frac{\|\mathbf{A}\|_1 \|\mathbf{W}\mathbf{A}^+\|_F}{\alpha\sqrt{\beta/2}}$, $\mathbf{A}$-strategy mechanism achieves (α, β)-WCQ accuracy.*

Proof. The noise vector added to the final query answer of $q_W(\cdot)$ using $\mathbf{A}$-strategy mechanism is $\hat{\boldsymbol{\eta}} = [\hat{\eta}_1, \ldots, \hat{\eta}_L] = (\mathbf{W}\mathbf{A}^+)\boldsymbol{\eta}$. Each noise variable $\hat{\eta}_i$ has a mean of 0, and a variance of $\sigma_i^2 = c_i \cdot (2b^2)$, where $c_i = \sum_{j=1}^{l} (\mathbf{W}\mathbf{A}^+)[i,j]^2$, and hence $\Pr[|\hat{\eta}_i| \geq \alpha] \leq \frac{2c_i b_{\mathbf{A}}^2}{\alpha^2}$ by Chebyshev's inequality. By union bound, the failing probability is bounded by

$$\begin{aligned} & \Pr[\|LM(\mathbf{W}, \mathbf{x}) - q_W(D)\|_\infty \geq \alpha] \\ = \;& \Pr[\cup_{i \in [1,L]} |\eta_i| \geq \alpha] \leq \sum_{i \in [1,L]} \frac{2c_i b^2}{\alpha^2} \leq \beta \end{aligned}$$

It requires $b \leq \frac{\alpha\sqrt{\beta/2}}{\|\mathbf{W}\mathbf{A}^+\|_F}$, hence $\epsilon \geq \|A\|_1/b \geq \frac{\|\mathbf{A}\|_1 \|\mathbf{W}\mathbf{A}^+\|_F}{\alpha\sqrt{\beta/2}}$. □

Proof for Theorem 5.3

Proof. Given an ϵ, the simulation in the function estimateFailingRateMC() ensures that with high probability $1-p$, the true failing probability β_t to bound $\|(\mathbf{W}\mathbf{A}^+)\eta\|_\infty$ by α lies within $\beta_e \pm \delta\beta$. The failing probability to bound $\beta_t < \beta_e + \delta\beta$ is $p/2$. By union bound, ϵ ensures (α, β')-WCQ accuracy, where $\beta' < \beta + \delta\beta + p/2$. If $(\beta + \delta\beta)(1-p/2) + p/2 < \beta + \delta\beta + p/2 < \beta$, then this ϵ ensures (α, β)-WCQ accuracy. Beside this estimation, in the binary search of translate(), we stop when $\epsilon_{\min}$ and $\epsilon_{\max}$ are sufficiently close. Hence, the privacy cost returned by translate() is an approximated minimal privacy cost required for (α, β)-WCQ accuracy. □

A.3 Multi-Poking Mechanism for ICQ

We would like to show Theorem 5.5 that given a query $q_{W,>c}$, multi-poking mechanism (Algorithm 4), achieves (α, β)-ICQ accuracy by executing function run$(q_{W,>c}, \alpha, \beta, D)$, with differential privacy cost of ϵ returned by function translate$(q_{W,>c}, \alpha, \beta)$.

Proof. For each $\phi \in W$, (i) when $q_\phi(D) < c - \alpha$,

$$\begin{aligned} & \Pr[\phi \in MPM^{\alpha,\beta}_{q_{\phi,>c}}(D) \mid c_\phi(D) < c - \alpha] \\ = \;& \sum_{i=0}^{m-1} \Pr[c_\phi(D) - c + \eta - \alpha_i + \alpha > 0 \mid c_\phi(D) - c + \alpha < 0] \\ < \;& \sum_{i=0}^{m-1} \Pr[\eta_i > \alpha_i] = \sum_{i=0}^{m-1} e^{-\alpha_i \epsilon_i / \|\mathbf{W}\|_1}/2 = m \cdot \beta/(mL) = \beta/L \end{aligned}$$

and (ii) when $q_\phi(D) < c + \alpha$,

$$\begin{aligned} & \Pr[\phi \notin MPM^{\alpha,\beta}_{\phi,>c}(D) \mid c_\phi(D) > c + \alpha] \\ = \;& \sum_{i=0}^{m-1} \Pr[c_\phi(D) - c + \eta + \alpha_i - \alpha < 0 \mid c_\phi(D) - c - \alpha > 0] \\ < \;& \sum_{i=0}^{m-1} \Pr[\eta_i < -\alpha_i] = \sum_{i=0}^{m-1} e^{-\alpha_i \epsilon_i / \|\mathbf{W}\|_1}/2 = m \cdot \beta/(mL) = \beta/L \end{aligned}$$

As $|W| = L$, the failing probability is bounded by β.

The RelaxPrivacy Algorithm [22] correlates the new noise η_{i+1} with noise η_i from the previous iteration In this way, the composition of the first $i+1$ iterations is ϵ_{i+1}, and the noise added in the $i+1$th iteration is equivalent to a noise generated with Laplace distribution with privacy budget ϵ_{i+1} and the first i iterations also satisfy ϵ_i-DP for $i = 0, 1, \ldots, m-1$ (Theorem 9 for single-dimension and Theorem 10 for high-dimension [22]). □

A.4 Laplace Top-k Mechanism for TCQ

We would like show the privacy and tolerance requirement of Laplace Top-k Mechanism stated in Theorem 5.6. The proof for accuracy is similar to the accuracy proof of Laplace mechanism for TCQ (Appendix A.1). Then we show that Algorithm 5 achieves minimal ϵ-differential privacy cost, where $\epsilon = k/b = \frac{2k\ln(L/(2\beta))}{\alpha}$.

Proof. Fix $D = D' \cup \{t\}$. Let $(x_1, \ldots, x_L)$, respectively $(x_1', \ldots, x_L')$, denote the vector of answers to the set of linear counting queries $q_{\phi_1}, \ldots, q_{\phi_L}$ when the table is D, respectively D'. Two properties are used: (1) Monotonicity of Counts: for all $j \in [L]$, $x_j \geq x_j'$; and (2) Lipschitz Property: for all $j \in [L]$, $1 + x_j' \geq x_j$.

Fix any k different values $(i_1, \ldots, i_k)$ from $[L]$, and fix noises $(\eta_{i_{k+1}}, \ldots, \eta_{i_L})$ drawn from $Lap(k/\epsilon)^{L-k}$ used for the remaining linear counting queries. Given these fixed noises, for $l \in \{i_1, \ldots, i_k\}$, we define

$$\eta_l^* = \min_{\eta} : (x_l + \eta > (\max_{j \in i_{k+1}, \ldots, i_L} x_j + \eta_j)) \tag{8}$$

For each $l \in \{i_1, \ldots, i_k\}$, we have

$$\begin{aligned} x_l' + (1 + \eta_l^*) &= (1 + x_l') + \eta_l^* \geq x_l + \eta_l^* \\ &> \max_{j \in i_{k+1}, \ldots, i_L} x_j + \eta_j \geq \max_{j \in i_{k+1}, \ldots, i_L} x_j' + \eta_j \end{aligned}$$

Hence, if $\eta_l \geq r_l^* + 1$ for all $l \in \{i_1, \ldots, i_k\}$, then $(i_1, \ldots, i_k)$ will be the output when the table is D' and the noise vector is $(\eta_{i_1}, \ldots, \eta_{i_k}, \ldots, \eta_L)$. The probabilities below are over the choices of $(\eta_{i_1}, \ldots, \eta_{i_k}) \sim Lap(k/\epsilon)^k$.

$$\begin{aligned} &\Pr[(i_1, \ldots, i_k) \mid D', \eta_{i_{i+1}}, \ldots, \eta_{i_L}] \\ \geq &\prod_{l \in \{i_1, \ldots, i_k\}} \Pr[\eta_l \geq 1 + \eta_l^*] \geq \prod_{l \in \{i_1, \ldots, i_k\}} e^{-k/\epsilon} \Pr[\eta_l \geq \eta_l^*] \\ \geq &\ e^{-\epsilon} \Pr[(i_1, \ldots, i_k) \mid D, \eta_{i_{i+1}}, \ldots, \eta_{i_L}] \end{aligned} \tag{9}$$

Proof of the other direction follows analogously.

Therefore, ϵ-differential privacy is guaranteed. □

B COMPOSITION THEOREMS

The sequential composition theorem helps assess the privacy loss of multiple differentialy private mechanisms.

Theorem B.1 (Sequential Composition [10]). *Let $M_1(\cdot)$ and $M_2(\cdot, \cdot)$ be algorithms with independent sources of randomness that ensure ϵ_1- and ϵ_2-differential privacy respectively. An algorithm that outputs both $M_1(D) = O_1$ and $M_2(O_1, D) = O_2$ ensures $(\epsilon_1 + \epsilon_2)$-differential privacy.*

Another important property of differential privacy is that postprocessing the outputs does not degrade privacy.

Theorem B.2 (Post-Processing [10]). *Let $M_1(\cdot)$ be an algorithm that satisfies ϵ-differential privacy. If applying an algorithm M_2 the output of $M_1(\cdot)$, then the overall mechanism $M_2 \circ M_1(\cdot)$ also satisfies ϵ-differential privacy.*

All steps in the post-processing algorithm do not access the raw data, so they do not affect the privacy analysis. Hence, any smart choices made by the data analyst after receiving the noisy answers from the privacy engine are considered as post-processing steps and do not affect the privacy analysis.

C ADDITIONAL EVALUATION

To further understand the relation between our accuracy requirement with commonly used error metric, we use F1 score to measure the similarity between the correct answer set and noisy answer set outputted from the mechanism. Figure 6 presents the F1 score of two queries: QI4 of an ICQ, and QT1 of a TCQ. In QT1, when α from $0.02|D|$ to $0.64|D|$, the median F1 score decreases from 0.9 to 0.15, which has a steeper gradient than the changes in Figure 2 and a closer trend with the theoretical α. In QI4, the F1 score is even more consistent with α and the empirical error shown in Figure 2. This shows that our (α, β) accuracy requirement is still a good indicator in data exploration process.

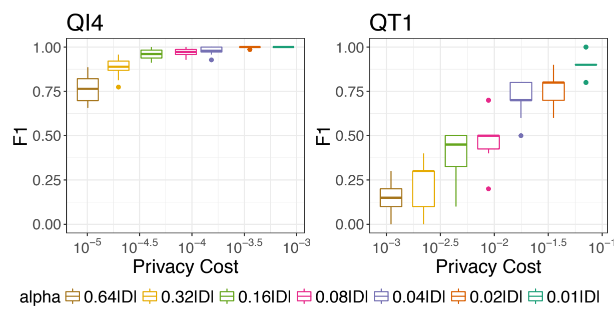

Figure 6: Privacy cost and empirical accuracy (F1 score) using optimal mechanism chosen by APEx (Optimistic Mode) on QI4 and QT1.

D RELATED WORK

DP in practice. Differential privacy [9] has emerged as a popular standard in real-world analysis products [11, 13, 14, 17, 25, 30]. There are well known techniques for special tasks such as answering linear counting queries[16, 27, 31], top-k queries [24] and frequent itemset mining [28, 36]. However, all the proposed work focuses on a particular type of query and there is no clear winner between these proposed algorithms for a given query. *A*PEx considers the key state-of-the-art differentially private algorithms [8, 25, 27] and makes decision on the algorithms on behalf of the data analyst.

General purpose frameworks like PINQ [32], wPINQ [33] and ϵktelo [37] allow users to write programs in higher level languages for various tasks. These systems automatically prove that every program written in this framework ensures differential privacy. FLEX [18] allows users to answer a SQL query under a specified privacy budget. Unlike all these systems, *A*PEx is the first that allows analysts ask a sequence of queries in high level language and specify only an accuracy bound, rather than a privacy level. *A*PEx automatically computes the privacy level to match the accuracy level.

Accuracy constraints in DP. Our problem is similar in spirit to Ligett et al. [29], which also considers analysts who specify accuracy constraints. Rather than concentrating on private machine learning theoretically, our focus is to explore private data with accuracy guarantees. The main technical differences are highlighted below.

First, the end-to-end problems in *A*PEx and the approach in Ligett et al. [29] are different. *A*PEx aims to translate a *given query with accuracy bound*, (q, α, β), to a differentially private mechanism that achieves this accuracy bound with the minimal privacy cost, (M, ϵ). On the other hand, Ligett et al. [29] takes as input a *given mechanism with accuracy bound and a set of privacy budgets*, $(M, \alpha, \beta, \{\epsilon_1 < \epsilon_2 < \cdots < \epsilon_T\})$, and outputs the minimal privacy cost (epsilon) for M to achieve the accuracy bound. Unlike Ligett et al., *A*PEx does not need to take a set of privacy budgets as an input. For the exploration queries in *A*PEx, there are more than just one differentially private mechanisms for each query type, and none of these mechanisms dominate the others. Thus, in this sense, the problem solved by *A*PEx is more general than the one solved by Ligett et al.

The second key difference between the approaches is the following. *A*PEx currently only supports mechanisms for which the relationship between the accuracy bound and the privacy loss epsilon can be established analytically. On the other hand, Ligett et al. can handle arbitrarily complex mechanisms, and use empirical error of the mechanisms to pick the epsilon. In this way, the solution proposed by Ligett et al. applies a larger class of mechanisms. Nevertheless, for the queries and mechanisms supported by *A*PEx, using Ligett et al.'s methods would be overkill in two ways: (1) there is an extra privacy cost ϵ_0 to test the empirical error, (2) since the exploration queries are variations of counts, the sensitivity of the error will be so high that the noise introduced to the empirical error could limit our ability to distinguish between different epsilon values. For example, for a simple counting query of size 1, the maximum privacy cost required in *A*PEx is $\ln(1/\beta)/\alpha$ while the privacy cost with Ligett et al. is more than $\epsilon_0 = 16(\ln(2T/\beta))/\alpha$. When *A*PEx applies a data-dependent approach, the privacy cost can be even smaller than $\ln(1/\beta)/\alpha$. For a prefix counting query of size L and sensitivity of L, the best privacy cost achieved in *A*PEx is $O(\log L)$, but ϵ_0 in [22] is $O(L)$ as the sensitivity of the error to the final query answer is L (for both Laplace and strategy-based mechanisms). Similarly, for iceberg counting queries (ICQ), the sensitivity of error to the final query answer is large and results in a large ϵ_0 for the differentially private testing. Thus, using the method in Ligett et al. would not help the currently supported mechanisms and queries in *A*PEx. In the future, we will add more complex mechanisms into *A*PEx (like DAWA [25], MWEM [15] that add data dependent noise) and study whether the methods of Ligett et al. can be adapted to our setting.

Data cleaning on private data. In our case study, we use *A*PEx as a means to help tune a cleaning workflow (namely entity resolution) on private data. Amongst prior work on cleaning private data [5, 23], the most relevant is PrivateClean [23] as it uses differential privacy too. However, similarities to *A*PEx stop with that. First, PrivateClean assumes a different setting, where no active data cleaner is involved. The data owner perturbs the dirty data without cleaning it, while the data analyst who wants to obtain some aggregate queries will clean the perturbed dirty data. Moreover, all the privacy perturbation techniques in PrivateClean are based on record-level perturbation, which (1) only work well for attributes with small domain sizes, and (2) has asymptotically poorer utility for aggregated queries. At last, the randomized response technique used for sampling string attributes in PrivateClean does not satisfy differential privacy – it leaks the active domain of the attribute.

Relationship to Privacy Laws. *A*PEx requires the data owner to only specify the overall privacy budget B. This setup may support and realize privacy laws in the future (e.g. GDPR [2]) to manage the privacy loss of users where B can be seen as the maximum budget to be spent without asking users' consent.

E QUERIES SUPPORTED BY *A*PEX

Besides supporting linear counting queries, our algorithms in *A*PEx can be easily extended to support SUM() queries. Queries for MEDIAN() (and percentile) can be supported by first querying the CDF (using a WCQ), and finding the median from that. GROUPBY can be expressed as a sequence of two queries: first query for values of an attribute with COUNT(*) > 0 (using ICQ), and then query the noisy counts for these bins (using WCQ). Similarly, if the aggregated function $f()$ in HAVING differs from the aggregated function $g()$, *A*PEx can express it as a sequence of two queries: first query for bins having $f() > c$ (using ICQ), and then apply $g()$ on these bins (using WCQ).

However, no differentially private algorithms can accurately report MAX() and MIN(). Non-linear queries such as AVG() and STD() can have very sensitive error bounds to noise when computed on a small number of rows (like the measures discussed in Section 3.2). Hence, supporting non-linear queries would need new translation mechanisms. Moreover, *A*PEx can support queries with foreign key joins (that do not amplify the contribution of one record), but designing accurate differentially private algorithms for queries with general joins is still an active area of research.