A Minimax Distortion View of Differentially Private Query Release

Weina Wang, Lei Ying and Junshan Zhang
School of Electrical, Computer and Energy Engineering
Arizona State University
{weina.wang, lei.ying.2, junshan.zhang}@asu.edu

***Abstract*—We devise query-set independent mechanisms for the problem of differentially private query release. Specifically, a differentially private mechanism is constructed to publish a synthetic database, and "customized" companion estimators are then derived to provide the best possible answers. Accordingly, the distortion corresponding to the best mechanism at the worst-case query, named the minimax distortion, provides a fundamental characterization. For the general class of statistical queries, by deriving asymptotically sharp upper and lower bounds, we prove that the minimax distortion is $O(1/n)$ as the database size n goes to infinity, with the squared-error distortion measure and fixed dimension of data entries.**

I. Introduction

It is envisaged that in the forthcoming "big data" era, there will be an abundance of rich data about individuals in many domains, such as healthcare, mobile networks, social networks and web search. While data analysis uncovers scientific and societal insights, it also poses potential "threats" to personal privacy. It is therefore of great interest to establish a systematic understanding of privacy-preserving data analysis, aiming to provide utility for data analytics while preserving privacy.

To rigorously quantify privacy, the celebrated notion of differential privacy, introduced in a line of work [1]–[3], has emerged as an analytical foundation for privacy-preserving data analysis. An information releasing mechanism is said to be *ϵ-differentially private* if the change of an individual's data alters the probability of any output instance by at most an e^{ϵ} multiplicative factor. By this requirement, the presence of the record associated with an individual, or the record's content, cannot be exactly deduced from the released information.

A central problem in differential privacy is how to provide accurate answers to as many as possible queries privately, which has been extensively studied by the literature through both interactive and non-interactive approaches (see, e.g., [1], [4]–[12]). Under the interactive approach, queries arrive online and each query consumes some privacy budget, and thus a delicate privacy allocation plan is needed. By contrast, the non-interactive approach uses all the privacy budget to generate a sanitized version of the database.

In this paper, we explore a non-interactive approach where a synthetic database is released by a differentially private mechanism whose form is independent of pre-given queries, which differentiates our work from the existing work [1], [4]–[11]. After the synthetic database is released, queries are answered by a "customized" estimator, rather than directly carried out as if the synthetic database were the actual database. In particular, the mechanism is constructed to "encode" rich stochastic structure into the synthetic database, and the estimator makes use of this structure (which is public information) and the query function. This approach decouples synthetic database generating and query answering. By introducing the flexibility of "customizing" estimators for different queries, it opens the possibility of deriving accurate answers for all queries in a general query class from the same released synthetic database.

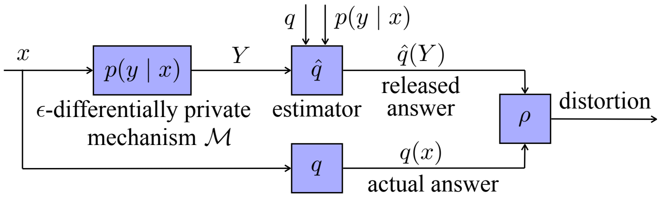

Fig. 1: Road map of our approach for differentially private query release.

Along this line, we take a minimax distortion view of differentially private query release. Consider a database consisting of n rows/entries, with each row having l binary attributes. Let the database be represented by a vector x of length n. Consider an ϵ-differentially private mechanism $\mathcal{M}$ for synthetic database release and let $Y = \mathcal{M}(x)$ denote the output. For each query q in a query class $\mathcal{Q}$, an estimator $\hat{q}$ is used to answer the query based on the synthetic database, and the answer is denoted by $\hat{q}(Y)$, as illustrated in Fig. 1. The accuracy of $\mathcal{M}$ for a query $q \in \mathcal{Q}$ is evaluated when an optimal estimator $\hat{q}^*$ is in use, since an optimal estimator fully exploits the available information in the mechanism. To guarantee accuracy for all queries in the query class, the performance of $\mathcal{M}$ is measured by the worst-case distortion among queries in $\mathcal{Q}$. Then a fundamental characterization of differentially private query release is the following minimax distortion:

$$\mathfrak{D}_{\epsilon} = \inf_{\substack{\epsilon\text{-differentially} \\ \text{private mechanisms}}} \sup_{q \in \mathcal{Q}, x \in \mathcal{D}^n} \mathbb{E}[\rho(\hat{q}^*(Y), q(x))], \quad (1)$$

where ρ is a distortion measure, $\hat{q}^*$ is the optimal estimator, and Y follows the probability distribution induced by x through the mechanism. This minimax distortion characterizes the best one can get from an ϵ-differentially private synthetic database releasing mechanism for the worst-case query accuracy guarantee, yielding a minimax distortion view of differentially private query release. Our main contributions are summarized as follows.

1) We propose a two-phase approach for differentially private query release: First, a synthetic database is released by a query-set independent differentially private mechanism, aiming at providing accurate answers for all queries in a general query class; then queries are answered by customized estimators. Based on this approach, we take a minimax distortion view of differentially private query release, where the minimax distortion $\mathfrak{D}_\epsilon$ is defined to be the distortion under the best ϵ-differentially private synthetic database releasing mechanism for the worst-case query in a general query class. Accordingly, the best mechanism enables all queries in a general query class to be answered with the associated distortion upper bounded by the minimax distortion.

2) For the class of statistical queries (which is a generalization of the class of linear queries in the literature), we consider the minimax distortion $\mathfrak{D}_\epsilon^{\mathrm{S}}$ with the squared-error distortion measure ρ, i.e., $\rho(s,t) = (s-t)^2$ for any $s, t \in \mathbb{R}$. We prove that the minimax distortion $\mathfrak{D}_\epsilon^{\mathrm{S}}$ is $O(1/n)$ by deriving asymptotically sharp upper and lower bounds in the regime that the database size n goes to infinity, for given data universe dimension l and privacy level ϵ.

The upper bound on $\mathfrak{D}_\epsilon^{\mathrm{S}}$ is achieved by a differentially private synthetic database releasing mechanism $\mathcal{E}$ and the companion estimators. The mechanism $\mathcal{E}$ can be viewed as an instance of the exponential mechanism and the randomized response mechanism. It encodes an independence structure into the released synthetic database that is exploited by the companion estimators. Under $\mathcal{E}$ and the estimators, all the statistical queries can be answered with distortion $O(1/n)$, which guarantees reasonable accuracy in large databases. The mechanism $\mathcal{E}$ satisfies the local model of differential privacy (see, e.g., [13]). However, we remark that the minimax distortion is for all differentially private mechanisms. We do not start from a local model but a local mechanism happens to be optimal in order.

Related Work: Differential privacy, introduced in the seminal work [1]–[3], has attracted extensive research studies.

Non-interactive approaches have been preferred by the data-mining community and the statistics community. However, some negative results have been found about this approach. Dinur and Nissim [14] showed that noise of magnitude $o(\sqrt{n})$ is blatantly non-private against $n \log^2 n$ random queries. Dwork et al. [1] found little statistical difference between the distributions induced by two databases that have very different answers to the same query. These negative results motivate interactive approaches, where the number of queries was initially limited to a sublinear order of n [1]. Subsequent work [6], [8] developed mechanisms that allow exponential number of predicate/linear queries to be answered with distortion $O(\mathrm{polylog}(|\mathcal{Q}|)/n^{1/3})$ and $O((\log(|\mathcal{Q}|))^{1/2}/n^{1/2})$, respectively, where the latter is for (ϵ, δ)-differential privacy.

Non-interactive approach was revisited by Blum, Ligett and Roth [4], where the distortion for each predicate query is upper bounded by $O((\mathrm{VCDIM}(\mathcal{Q}))^{1/3}/n^{1/3})$, with $\mathrm{VCDIM}(\mathcal{Q})$ being the VC-dimension of a concept class $\mathcal{Q}$. A similar distortion bound $O((\log(|\mathcal{Q}|))^{1/3}/n^{1/3})$ was achieved by the work of Hardt, Ligett and McSherry [9] for linear queries. A distortion bound $O((\log(|\mathcal{Q}|))^{1/2}/n^{1/2})$ under (ϵ, δ)-differential privacy was also achieved in their work.

The minimax distortion studied in this paper is different from the statistical minimax risk, which is a classical framework for parameter estimation. Statistical minimax risk with constraint of local differential privacy has been studied by Duchi, Jordan and Wainwright [15], [16].

II. System Model

We consider the following model for a database. A database is represented by a vector x of length n, with each entry corresponding to a row of the database and n being the size of the database. Entries of x are denoted by $x_1, x_2, \ldots, x_n$, and they take values from a domain $\mathcal{D} = \{0,1\}^l$, i.e., they have l binary attributes. Then $\mathcal{D}^n = (\{0,1\}^l)^n$ denotes the set of all possible databases. Two databases $x, x' \in \mathcal{D}^n$ are said to be *neighbors* if they differ on exactly one row, and $x \sim x'$ denotes the neighboring relation.

Information about a database is acquired through queries. A *query* is a function $q\colon \mathcal{D}^n \to \mathcal{R}$, where $\mathcal{R}$ is some abstract range. Consider a database $x \in \mathcal{D}^n$. The answer $q(x)$ to the query contains information about x; however, directly releasing $q(x)$ may compromise privacy, necessitating privacy-preserving information releasing mechanisms.

Definition 1. A *mechanism* $\mathcal{M}$ is specified by an *associated mapping* $\mu_{\mathcal{M}}\colon \mathcal{D}^n \to \mathcal{P}$, where $\mathcal{P}$ is the set of probability measures on some measurable space $(\mathcal{S}, \mathcal{F})$, called the *range* of the mechanism $\mathcal{M}$. Taking a database $x \in \mathcal{D}^n$ as the input, the mechanism $\mathcal{M}$ outputs an $\mathcal{S}$-valued random variable with distribution measure $\mu_{\mathcal{M}}(x)$ on $(\mathcal{S}, \mathcal{F})$.

Definition 2. (Dwork et al. [1], [2]) A mechanism $\mathcal{M}$ is ϵ-*differentially private* for some $\epsilon \in [0, +\infty]$ if for any pair of neighboring databases $x, x' \in \mathcal{D}^n$, and any measurable $\mathcal{K} \in \mathcal{F}$,

$$\mathbb{P}\{\mathcal{M}(x) \in \mathcal{K}\} \leq e^\epsilon \mathbb{P}\{\mathcal{M}(x') \in \mathcal{K}\}. \tag{2}$$

Intuitively, differential privacy requires certain indistinguishability between the distributions induced by neighboring databases. The smaller ϵ is, the more indistinguishability is required, and hence the better privacy is. We call the parameter ϵ the *level of differential privacy*. Note that the differential privacy property of a mechanism is fully characterized by its associated mapping.

III. Minimax Distortion

We consider differentially private mechanisms for non-interactive synthetic database release. Specifically, let $\wp(\mathcal{D}^n)$ denote the power set of $\mathcal{D}^n$. Then we consider differentially private mechanisms with range $(\mathcal{D}^n, \wp(\mathcal{D}^n))$. Let $\mathcal{M}$ be such a mechanism and $x \in \mathcal{D}^n$ be a database. Then the output $Y = \mathcal{M}(x)$ is a $\mathcal{D}^n$-valued random variable, representing the released synthetic database. For each query q in a query class $\mathcal{Q}$, an estimator $\hat{q}\colon \mathcal{D}^n \to \mathcal{R}$ is used to answer the query based on the synthetic database, and thus the answer is denoted by $\hat{q}(Y)$. The distortion between the actual answer $q(x)$ and the released answer $\hat{q}(Y)$ is measured by a distortion measure ρ on the range of the query q. Note that as long as $\mathcal{M}$ is ϵ-differentially private, arbitrary number of queries can be answered and any estimator can be used, with the level of differential privacy still preserved.

The proposed approach aims at privately releasing a synthetic database that permits accurate answers to be derived for all queries in a general query class. Therefore, a natural fundamental characterization of differentially private query release is the following minimax distortion: the distortion under the best differentially private synthetic database releasing mechanism (the "min" part) for the worst-case query in the query class (the "max" part). In what follows, we derive the formal definition of the minimax distortion.

For the sake of fair comparison, we assume that q is normalized, i.e., $\max_{x,x'\in\mathcal{D}^n}\rho(q(x),q(x'))=1$, which rules out trivial queries that map all possible databases to a constant. For each query q, to guarantee that the released answers have "physical meanings," we consider the estimators such that the answers released by them correspond to possible answers to the query q on real databases, i.e., the estimators in $\hat{\mathcal{Q}}_q=\{\hat{q}\colon\mathcal{D}^n\to\mathcal{R}\mid\hat{q}(\mathcal{D}^n)\subseteq q(\mathcal{D}^n)\}$, which we call *proper estimators*. For an estimator $\hat{q}\in\hat{\mathcal{Q}}_q$ for the query q, let us consider the worst-case distortion among all possible databases, i.e., $\sup_{x\in\mathcal{D}^n}\mathbb{E}_{Y\sim\mu_{\mathcal{M}}(x)}[\rho(\hat{q}(Y),q(x))]$, where the subscript $Y\sim\mu_{\mathcal{M}}(x)$ indicates that Y follows the distribution $\mu_{\mathcal{M}}(x)$, and the expectation is taken over all the randomness.

To minimize distortion, an estimator should be designed according to the mechanism $\mathcal{M}$ and the query q, making use of all the available information, which is illustrated in Fig. 1. Therefore an optimal estimator $\hat{q}^*$ is given by

$$\hat{q}^*\in\arg\inf_{\hat{q}\in\hat{\mathcal{Q}}_q}\sup_{x\in\mathcal{D}^n}\mathbb{E}_{Y\sim\mu_{\mathcal{M}}(x)}[\rho(\hat{q}(Y),q(x))].\tag{3}$$

Note that the set $\hat{\mathcal{Q}}_q$ contains only a finite number of estimators since it consists of mappings from $\mathcal{D}^n$ to $q(\mathcal{D}^n)$, which are both finite sets, indicating that the infimum in (3) can be attained. Since the information in a mechanism is fully exploited only when an optimal estimator is in use, the accuracy of an ϵ-differentially private mechanism $\mathcal{M}$ for a query q is evaluated with an optimal estimator $\hat{q}^*$, i.e., by the distortion $\sup_{x\in\mathcal{D}^n}\mathbb{E}_{Y\sim\mu_{\mathcal{M}}(x)}[\rho(\hat{q}^*(Y),q(x))]$.

The synthetic database Y released by $\mathcal{M}$ is expected to answer all queries in a query class $\mathcal{Q}$. To guarantee accuracy for all queries in $\mathcal{Q}$, the performance of $\mathcal{M}$ is measured by the worst-case distortion among all queries in $\mathcal{Q}$, i.e., by

$$\sup_{q\in\mathcal{Q}}\left\{\sup_{x\in\mathcal{D}^n}\mathbb{E}_{Y\sim\mu_{\mathcal{M}}(x)}[\rho(\hat{q}^*(Y),q(x))]\right\}.\tag{4}$$

Let $\mathcal{U}_\epsilon$ be the set of mappings associated with ϵ-differentially private mechanisms. The *minimax distortion* is defined as

$$\mathfrak{D}_\epsilon=\inf_{\mu_{\mathcal{M}}\in\mathcal{U}_\epsilon}\sup_{q\in\mathcal{Q}}\left\{\sup_{x\in\mathcal{D}^n}\mathbb{E}_{Y\sim\mu_{\mathcal{M}}(x)}[\rho(\hat{q}^*(Y),q(x))]\right\}.\tag{5}$$

IV. Statistical Queries

In this section, we study the minimax distortion for the class of statistical queries.

Definition 3. A *statistical query* $q_\varphi\colon\mathcal{D}^n\to\mathbb{R}$ is specified by a sequence of functions

$$\varphi=(\varphi_i\colon\mathcal{D}\to\mathbb{R},i=1,2,\dots),\tag{6}$$

where each φ_i is a function of the ith row of the database, which we call a *row function*, and there is no constraint on its form except boundedness. Let $a_i=\min_{v\in\mathcal{D}}\varphi_i(v)$, $b_i=\max_{v\in\mathcal{D}}\varphi_i(v)$ and $c_i=b_i-a_i$. Assume that for any $i\in[n]$, $a\le a_i<b_i\le b$ and $c_i\ge c$ for some $a,b,c\in\mathbb{R}$ with $c>0$. Then q_φ is defined by

$$q_\varphi(x)=\frac{1}{\sum_{i=1}^n c_i}\sum_{i=1}^n\varphi_i(x_i),\tag{7}$$

where $x_1,\dots,x_n$ are the rows of the database x.

Note that the above definition of statistical query is a generalization of the so called *linear query* (and its special form *predicate/counting query*) in the literature [4], [6], [8]–[10], [18], [19], since a linear query can be written as a statistical query with identical row functions for all the rows. Linear queries can be answered as long as the histogram of a database is known. However, histograms are often not sufficient for answering statistical queries, making the approaches that privately release histograms not applicable for statistical queries.

Denote the class of statistical queries by $\mathcal{Q}^{\mathrm{S}}$ and consider the squared-error distortion measure ρ, i.e., $\rho(s,t)=(s-t)^2$ for any $s,t\in\mathbb{R}$. Then the minimax distortion for statistical queries can be written as

$$\mathfrak{D}^{\mathrm{S}}_\epsilon=\inf_{\mu_{\mathcal{M}}\in\mathcal{U}_\epsilon}\sup_{q_\varphi\in\mathcal{Q}^{\mathrm{S}},x\in\mathcal{D}^n}\mathbb{E}_{Y\sim\mu_{\mathcal{M}}(x)}\left[|\hat{q}^*_\varphi(Y)-q_\varphi(x)|^2\right].\tag{8}$$

Theorem 1. *The minimax distortion for statistical queries satisfies the following bounds:*

$$\frac{\big(1-\Phi(1)\big)^2}{2^{l+4}\big(1+\frac{e^\epsilon}{2^l-1}\big)^3}\frac{1}{n}+o\left(\frac{1}{n}\right)\le\mathfrak{D}^{\mathrm{S}}_\epsilon\le\frac{4(b-a)^2\big(1+(2^l-1)e^{-\epsilon}\big)^2}{c^2(1-e^{-\epsilon})^2}\frac{1}{n},\tag{9}$$

where Φ is the cumulative distribution function (CDF) of the standard Gaussian distribution, and a,b,c are the constants in Definition 3.

Consider the asymptotic regime that the database size n goes to infinity for given data universe dimension l and privacy level ϵ. Then the upper bound indicates that there exist query-set independent differentially private synthetic database releasing mechanisms and estimators such that all the statistical queries can be answered with distortion $O(1/n)$. Further, the lower bound and the upper bound are of the same order in terms of database size, which shows that these bounds are asymptotically tight in the considered regime. We derive these bounds in the following subsections.

Remark. We caution that when the privacy level ϵ or the data universe dimension l also scales, the upper and lower bounds given here may not meet. For example, let $\epsilon = n^{-\beta}$ for some $\beta > 0$ and consider the joint asymptotic regime on the 2-dimensional $(n, 1/\epsilon)$-plane. In this case, the upper and lower bounds differ by a factor of the order of $n^{2\beta}$.

A. Upper Bound on the Minimax Distortion

According to the definition of the minimax distortion, the distortion under some specific ϵ-differentially private mechanism and estimators serves as an upper bound on $\mathfrak{D}^{\mathrm{S}}_{\epsilon}$.

Consider a synthetic database releasing mechanism $\mathcal{E}$ with associated mapping $\mu_{\mathcal{E}}$. For each database $x \in \mathcal{D}^n$, we use the PMF $p_{\mathcal{E}(x)}$ to represent the distribution measure $\mu_{\mathcal{E}}(x)$ since the output $\mathcal{E}(x)$ has a discrete alphabet $\mathcal{D}^n$. Then let the mechanism $\mathcal{E}$ be specified by

$$p_{\mathcal{E}(x)}(y) = \frac{e^{-\epsilon d(x,y)}}{\big(1 + (2^l - 1)e^{-\epsilon}\big)^n}, \quad x, y \in \mathcal{D}^n, \tag{10}$$

where $\epsilon \in [0, +\infty)$ and d is the Hamming distance on $\mathcal{D}^n$. By the form of $p_{\mathcal{E}(x)}$, the mechanism $\mathcal{E}$ can be cast as an instance of the exponential mechanism with score function $-d$ [20].

Let Y denote $\mathcal{E}(x)$ and Y_i denote the ith row of Y. Then by (10), the entries $\{Y_i, i \in [n]\}$ are independent and each entry Y_i has the following PMF

$$p_{Y_i}(y_i) = \frac{e^{-\epsilon \delta(x_i, y_i)}}{1 + (2^l - 1)e^{-\epsilon}}, \quad y_i \in \mathcal{D}. \tag{11}$$

Therefore, the mechanism $\mathcal{E}$ can also be viewed as a randomized response scheme, where the released database is generated by perturbing each individual's data independently and distributedly.

The differential privacy property of $\mathcal{E}$ is given in the following lemma. The proof is standard and thus we omit it here due to space limit.

Lemma 1. *The mechanism $\mathcal{E}$ is ϵ-differentially private.*

Next we present the estimators companioned with $\mathcal{E}$ for the class of statistical queries. Let $g(\epsilon) = 1 + (2^l - 1)e^{-\epsilon}$. For each $q_\varphi \in \mathcal{Q}^{\mathrm{S}}$, consider the estimator $\hat{q}^{\mathrm{u}}_\varphi \colon \mathcal{D}^n \to \mathcal{R}$ defined by

$$\hat{q}^{\mathrm{u}}_\varphi(y) = \frac{g(\epsilon)}{1 - e^{-\epsilon}} q_\varphi(y) - \frac{e^{-\epsilon}}{1 - e^{-\epsilon}} C_\varphi, \tag{12}$$

where

$$C_\varphi = \frac{1}{\sum_{i=1}^n c_i} \sum_{i=1}^n \sum_{v \in \mathcal{D}} \varphi_i(v). \tag{13}$$

The answer given by $\hat{q}^{\mathrm{u}}_\varphi$ may not always be consistent with an actual database, in which case $\hat{q}^{\mathrm{u}}_\varphi \notin \hat{\mathcal{Q}}_{q_\varphi}$. Thus we consider the estimator $\hat{q}_\varphi \colon \mathcal{D}^n \to \mathcal{R}$ defined by

$$\hat{q}_\varphi(y) \in \operatorname*{arg\,min}_{r \in q_\varphi(\mathcal{D}^n)} |\hat{q}^{\mathrm{u}}_\varphi(y) - r|, \tag{14}$$

which quantizes the answer given by $\hat{q}^{\mathrm{u}}_\varphi$ to the closest value in $q_\varphi(\mathcal{D}^n)$ and thus guarantees that $\hat{q}_\varphi$ is a proper estimator.

Lemma 2. *Under the mechanism $\mathcal{E}$, the distortion of the estimator $\hat{q}_\varphi$ satisfies the following upper bound:*

$$\sup_{x \in \mathcal{D}^n} \mathbb{E}\big[|\hat{q}_\varphi(Y) - q_\varphi(x)|^2\big] \le \frac{4(b-a)^2\big(1 + (2^l - 1)e^{-\epsilon}\big)^2}{c^2(1 - e^{-\epsilon})^2} \frac{1}{n}, \tag{15}$$

where a, b, c are the constants in Definition 3.

The proof of this lemma is given in Appendix B. In the proof, we first quantify the distortion of $\hat{q}^{\mathrm{u}}_\varphi$, which is a quarter of the upper bound in (15). The intuition is that the mechanism $\mathcal{E}$ perturbs each row of the underlying database independently, which encodes an independence structure into the released synthetic base, and then the estimator $\hat{q}^{\mathrm{u}}_\varphi$ exploits this structure. By the law of large numbers (LLN), the aggregate perturbation converges to the expectation, which is a constant determined by the query and thus can be removed in the estimator. Then we show that the quantization in $\hat{q}_\varphi$ degrades the performance guarantee by a factor no greater than 4.

Compared with existing approaches, the synthetic database releasing mechanism $\mathcal{E}$ does not require a priori knowledge of the queries of interest, and instead of answering query q_φ by $q_\varphi(Y)$, the estimators $\hat{q}^{\mathrm{u}}_\varphi$ and $\hat{q}_\varphi$ make more use of the stochastic structure in Y encoded by the mechanism $\mathcal{E}$.

Remark. In many cases, the value C_φ in the estimator $\hat{q}^{\mathrm{u}}_\varphi$ can be easily obtained rather than exhaustive calculation. See Appendix C for an example. The estimator $\hat{q}^{\mathrm{u}}_\varphi$ is more computationally efficient than $\hat{q}_\varphi$ since it does not need to find the value closest to $\hat{q}^{\mathrm{u}}_\varphi(Y)$ in $q_\varphi(\mathcal{D}^n)$. Therefore, when we are not constricted to proper estimators, it is more desirable to use $\hat{q}^{\mathrm{u}}_\varphi$ from an implementation perspective.

B. Lower Bound on the Minimax Distortion

Consider any ϵ-differentially private mechanism $\mathcal{M}$. For any query $q_\varphi \in \mathcal{Q}^{\mathrm{S}}$, the form of the optimal estimator depends on q_φ. Therefore with slight abuse of notation, we denote the optimal estimator by the function $\hat{q}^* \colon \mathcal{D}^n \times \mathcal{Q}^{\mathrm{S}} \to \mathbb{R}$ and the answer by $\hat{q}^*(Y, q_\varphi)$, where Y is the synthetic database released by the mechanism $\mathcal{M}$. Then our goal is to derive a lower bound on the following worst-case distortion:

$$\sup_{q_\varphi \in \mathcal{Q}^{\mathrm{S}}, x \in \mathcal{D}^n} \mathbb{E}_{Y \sim \mu_{\mathcal{M}}(x)}\big[|\hat{q}^*(Y, q_\varphi) - q_\varphi(x)|^2\big]. \tag{16}$$

Consider such a type of queries, each of which is specified by an element $z \in \mathcal{D}^n$ and defined by $q_z(x) = \frac{1}{n} d(x, z)$ for any $x \in \mathcal{D}^n$, where d is the Hamming distance on $\mathcal{D}^n$. For any $v, v' \in \mathcal{D}$, let $\delta(v, v') = 0$ if $v = v'$ and $\delta(v, v') = 1$ otherwise. Then the query q_z can be written as $q_z(x) = \frac{1}{n} \sum_{i=1}^n \delta(x_i, z_i)$, which indicates that q_z is a statistical query. Let

$$\mathcal{Q}^{\mathrm{Z}} = \left\{ q_z \colon \mathcal{D}^n \to \mathbb{R} \;\middle|\; q_z(x) = \frac{1}{n} d(x, z), z \in \mathcal{D}^n \right\}. \tag{17}$$

Then $\mathcal{Q}^{\mathrm{Z}} \subseteq \mathcal{Q}^{\mathrm{S}}$, and therefore the supremum (16) is no smaller than the supremum over $q_z \in \mathcal{Q}^{\mathrm{Z}}$.

To derive a lower bound on the above supremum over $q_z \in \mathcal{Q}^{\mathrm{Z}}$, consider $\mathcal{D}^n$-valued random variables X, Y and Z, where X follows a uniform distribution. Given $X = x$, the conditional PMF of Y is specified by the distribution measure $\mu_{\mathcal{M}}(x)$, i.e., $p_{Y|X}(y \mid x) = \mathbb{P}\{\mathcal{M}(x) = y\}$ for any $y \in \mathcal{D}^n$. The random variable Z is independent of X and Y and also follows a uniform distribution.

Consider the query q_Z, which is the query in $\mathcal{Q}^{\mathrm{Z}}$ specified by the random variable Z. Then

$$\begin{aligned}
&\sup_{q_z \in \mathcal{Q}^{\mathrm{Z}}, x \in \mathcal{D}^n} \mathbb{E}_{Y \sim \mu_{\mathcal{M}}(x)}\big[|\hat{q}^*(Y, q_z) - q_z(x)|^2\big] \\
&\overset{(a)}{=} \sup_{q_z \in \mathcal{Q}^{\mathrm{Z}}, x \in \mathcal{D}^n} \mathbb{E}\big[|\hat{q}^*(Y, q_Z) - q_Z(X)|^2 \bigm| X = x, Z = z\big] \\
&\geq \sum_{z \in \mathcal{D}^n, x \in \mathcal{D}^n} \mathbb{E}\big[|\hat{q}^*(Y, q_Z) - q_Z(X)|^2 \bigm| X = x, Z = z\big] p_X(x) p_Z(z) \\
&= \mathbb{E}\big[|\hat{q}^*(Y, q_Z) - q_Z(X)|^2\big],
\end{aligned}$$

where (a) is due to the independence between Z and (X, Y). Note that we construct the random variables X and Z only for the proof. Our result in Theorem 1 does not assume any stochastic model for the database or the query. Note that $\hat{q}^*(Y, q_Z)$ is a function of Y and Z. Since the conditional expectation is precisely the minimum mean square estimator [21], we have

$$\begin{aligned}
&\mathbb{E}\big[|\hat{q}^*(Y, q_Z) - q_Z(X)|^2\big] \\
&\geq \mathbb{E}\big[|\mathbb{E}[q_Z(X) \mid Y, Z] - q_Z(X)|^2\big] \qquad (18) \\
&= \frac{1}{n^2} \mathbb{E}\big[|\mathbb{E}[d(X, Z) \mid Y, Z] - d(X, Z)|^2\big]. \qquad (19)
\end{aligned}$$

Recall that the conditional PMF $p_{Y|X}(\cdot \mid x)$ is specified by the distribution measure $\mu_{\mathcal{M}}(x)$. Then since the mechanism $\mathcal{M}$ is ϵ-differentially private, for any neighboring $x, x' \in \mathcal{D}^n$ and any $y \in \mathcal{D}^n$, $p_{Y|X}(y \mid x) \leq e^{\epsilon} p_{Y|X}(y \mid x')$. This inequality is needed in the proof of the following lemma, which gives a lower bound on the expectation in (19).

Lemma 3. *There exists a constant C such that*

$$\begin{aligned}
&\mathbb{E}\big[|\mathbb{E}[d(X, Z) \mid Y, Z] - d(X, Z)|^2\big] \\
&\geq \frac{1}{4}\bigg(\big(1 - \Phi(1)\big)\sigma\gamma^{\frac{3}{2}}\sqrt{n} - \frac{C\rho\gamma}{\sigma^3}\bigg)^2, \qquad (20)
\end{aligned}$$

where Φ is the CDF of the standard Gaussian distribution,

$$\gamma = \frac{1}{2\big(1 + \frac{e^{\epsilon}}{2^l - 1}\big)}, \quad \sigma^2 = \frac{1}{2^{l-1}}, \quad \rho = \frac{1}{2^{l-1}}. \qquad (21)$$

The proof is presented in Appendix D. By this lemma, for any ϵ-differentially private mechanism $\mathcal{M}$, the distortion is lower bounded as

$$\begin{aligned}
&\sup_{q_\varphi \in \mathcal{Q}^{\mathrm{S}}, x \in \mathcal{D}^n} \mathbb{E}_{Y \sim \mu_{\mathcal{M}}(x)}\big[|\hat{q}^*(Y, q_\varphi) - q_\varphi(x)|^2\big] \\
&\geq \frac{\big(1 - \Phi(1)\big)^2}{2^{l+4}\big(1 + \frac{e^{\epsilon}}{2^l - 1}\big)^3} \frac{1}{n} + o\bigg(\frac{1}{n}\bigg), \qquad (22)
\end{aligned}$$

which further implies the lower bound in Theorem 1.

V. Conclusion and Future Work

A two-phase approach was proposed for differentially private query release, where a fundamental characterization was given in terms of the minimax distortion. For the general class of statistical queries, asymptotically sharp bounds on the minimax distortion were derived in the regime that the database size n goes to infinity. Our future research interest includes the joint asymptotic regime in terms of n, the data universe dimension l and the differential privacy level ϵ.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Conf. Theory of Cryptography (TCC)*, New York, NY, 2006, pp. 265–284.

[2] C. Dwork, "Differential privacy," in *Proc. Int. Conf. Automata, Languages and Programming (ICALP)*, Venice, Italy, 2006, pp. 1–12.

[3] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, St. Petersburg, Russia, 2006, pp. 486–503.

[4] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to non-interactive database privacy," in *Proc. Ann. ACM Symp. Theory of Computing (STOC)*, Victoria, Canada, 2008, pp. 609–618.

[5] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan, "On the complexity of differentially private data release: efficient algorithms and hardness results," in *Proc. Ann. ACM Symp. Theory of Computing (STOC)*, Bethesda, MD, 2009, pp. 381–390.

[6] A. Roth and T. Roughgarden, "Interactive privacy via the median mechanism," in *Proc. Ann. ACM Symp. Theory of Computing (STOC)*, Cambridge, MA, 2010, pp. 765–774.

[7] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *Proc. Ann. IEEE Symp. Found. Comput. Sci. (FOCS)*, Las Vegas, NV, 2010, pp. 51–60.

[8] M. Hardt and G. N. Rothblum, "A multiplicative weights mechanism for privacy-preserving data analysis," in *Proc. Ann. IEEE Symp. Found. Comput. Sci. (FOCS)*, Las Vegas, NV, 2010, pp. 61–70.

[9] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," in *Advances Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, Dec. 2012, pp. 2348–2356.

[10] A. Gupta, A. Roth, and J. Ullman, "Iterative constructions and private data release," in *Proc. Conf. Theory of Cryptography (TCC)*, Sicily, Italy, 2012, pp. 339–356.

[11] M. Gaboardi, E. J. G. Arias, J. Hsu, A. Roth, and Z. S. Wu, "Dual query: Practical private query release for high dimensional data," in *Int. Conf. Machine Learning (ICML)*, Beijing, China, 2014.

[12] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy, and mutual-information privacy," in *Proc. Ann. Allerton Conf. Commununication, Control and Computing*, Monticello, IL, Sep. 2014, pp. 1086–1092.

[13] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, May 2011.

[14] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Symp. Principles Database Systems (PODS)*, San Diego, CA, 2003, pp. 202–210.

[15] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and minimax bounds: Sharp rates for probability estimation," in *Advances Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, Dec. 2013, pp. 1529–1537.

[16] ——, "Local privacy and statistical minimax rates," in *Proc. Ann. IEEE Symp. Found. Comput. Sci. (FOCS)*, Berkeley, CA, Oct. 2013, pp. 429–438.

[17] M. Hardt and K. Talwar, "On the geometry of differential privacy," in *Proc. Ann. ACM Symp. Theory of Computing (STOC)*, Cambridge, MA, 2010, pp. 705–714.

[18] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor, "Optimizing linear counting queries under differential privacy," in *Symp. Principles Database Systems (PODS)*, Indianapolis, IN, 2010, pp. 123–134.

[19] J. Ullman, "Answering $n^{2+o(1)}$ counting queries with differential privacy is hard," in *Proc. Ann. ACM Symp. Theory of Computing (STOC)*, Palo Alto, CA, 2013, pp. 361–370.
[20] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. Ann. IEEE Symp. Found. Comput. Sci. (FOCS)*, Providence, RI, 2007, pp. 94–103.
[21] K. B. Athreya and S. N. Lahiri, *Measure Theory and Probability Theory*. New York, NY: Springer, 2006.
[22] "Netflix Prize," http://www.netflixprize.com.
[23] J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Advances Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, 2012, pp. 548–556.
[24] J. Blocki, A. Blum, A. Datta, and O. Sheffet, "The johnson-lindenstrauss transform itself preserves differential privacy," in *Proc. Ann. IEEE Symp. Found. Comput. Sci. (FOCS)*, New Brunswick, NJ, 2012, pp. 410–419.
[25] A. Gupta, M. Hardt, A. Roth, and J. Ullman, "Privately releasing conjunctions and the statistical query barrier," in *Proc. Ann. ACM Symp. Theory of Computing (STOC)*, San Jose, CA, 2011, pp. 803–812.
[26] K. L. Chung, *A Course in Probability Theory*, 3rd ed. San Diego, CA: Academic Press, 2000.

Appendix A
Experimental Evaluation and Application

In this section, we first evaluate the mechanism $\mathcal{E}$ in (10) when companioned with the estimator $\hat{q}^{\mathrm{u}}_{\varphi}$ in (12) through experiments on a Netflix dataset [22] for statistical queries. During the experiments, we compare our approach with the MWEM algorithm (a combination of the Exponential Mechanism with the Multiplicative Weights update rules) [9]. The main conclusion from the experimental results is that the proposed approach provides reasonable accuracy for all the tested queries, irrespective of the form of the queries or the number of the tested queries, which improves over the MWEM algorithm. The scaling behavior $O(1/n)$ of the minimax distortion as the database size n goes to infinity is also verified by the experimental results.

We next consider the application of differentially private cut function release for graphs and derive an upper bound on the minimax distortion for this application. We evaluate our approach through experiments on a Facebook dataset [23]. The experimental results verify the theoretical upper bound and show that the proposed approach works well for this application.

A. Evaluation for Statistical Queries

In this subsection, we conduct experiments on the Netflix dataset for statistical queries. The Netflix dataset consists of movie ratings from users, with each rating on a scale from 1 to 5 (integral) stars. We treat each rating as a row and model the dataset as a database. To obtain databases with different sizes, we take subsets from the dataset.

The experimental evaluation in this subsection has three focuses: (1) the separation between statistical queries and linear queries, (2) distortion under varying query set size, and (3) scaling behavior of the distortion under varying database size.

Note that for the convenience of comparison, we often calculate the absolute-error distortion when evaluating the mechanism $\mathcal{E}$ and the companion estimator $\hat{q}^{\mathrm{u}}_{\varphi}$. Under the absolute-error distortion defined by $\rho(s,t) = |s-t|$, for any $s,t \in \mathbb{R}$, the distortion upper bound for the estimator $\hat{q}^{\mathrm{u}}_{\varphi}$ becomes

$$\sup_{x\in\mathcal{D}^n} \mathbb{E}\big[|\hat{q}^{\mathrm{u}}_{\varphi}(Y) - q_{\varphi}(x)|\big] \leq \frac{(b-a)\big(1+(2^l-1)e^{-\epsilon}\big)}{c(1-e^{-\epsilon})}\frac{1}{\sqrt{n}},$$

since by Jensen's inequality $\big(\mathbb{E}\big[|X|\big]\big)^2 \leq \mathbb{E}\big[|X|^2\big]$ for any random variable X.

1) Statistical Queries vs. Linear Queries: The class of statistical queries is much larger than the class of linear queries since a statistical query allows different row functions, whereas a linear query can only have identical row functions. For the private movie rating release application, it is possible to encounter queries that perform different functions on different rows, since different movies or users may belong to different groups and have different weights in a query. We call the number of distinct row functions in a statistical query the *heterogeneity* of the query. Consider a statistical query q_φ and the associated row function sequence $\varphi = (\varphi_1, \dots, \varphi_n)$. If the heterogeneity of q_φ equals to 1, then $\varphi_1 = \cdots = \varphi_n$, and thus q_φ is a linear query. If the heterogeneity of q_φ is greater than 1, then not all the φ_i's are equal. For example, during the experiments in this subsection, when the heterogeneity equal to 2, the statistical query performs one row function for the first half of the rows, and performs another row function for the second half.

The mechanism $\mathcal{E}$ and the companion estimator $\hat{q}^{\mathrm{u}}_{\varphi}$ is designed for statistical queries. The upper bound on the distortion of the proposed approach holds for any statistical query, and thus holds for any heterogeneity. The MWEM algorithm is designed for linear queries. To evaluate the MWEM algorithm for statistical queries, we adapt it as follows. For each distinct row function in a statistical query, we treat the set of rows associated with this row function as a "sub-database". Restricted to this sub-database, the statistical query is a linear query, so we can run the MWEM algorithm on the sub-database to generate a synthetic sub-database. Then the answer to the statistical query is obtained by combining the answers at each sub-database. When there are multiple statistical queries in the query set, we need to divide the database into sub-databases such that restricted to a sub-database, any query in the query set is a linear query. In the experiment, we consider statistical queries with same row functions for ratings of the same movie.

We evaluate the proposed approach and the MWEM algorithm on a database of size $n = 162,567$ from the Netflix dataset, consisting of ratings for 128 movies. Each movie has roughly $1000 \sim 2000$ ratings. Statistical queries are generated randomly in the following way. To specify a row function φ_i, the values $\varphi(1), \varphi(2), \dots, \varphi(5)$ are sufficient. We generate i.i.d. random variables $X_1, \dots, X_5$ with uniform distribution on $[0,1]$, and divide them by $\max_i X_i - \min_i X_i$ for normalization. Then these values are used to specify a row function. For a statistical query with heterogeneity h, we generate h row functions independently, and assign each row function to rows corresponding to $1/h$ of the movies. During the experiments, we consider heterogeneity varying from 1 to 128. For each heterogeneity h, we generate a set of 200

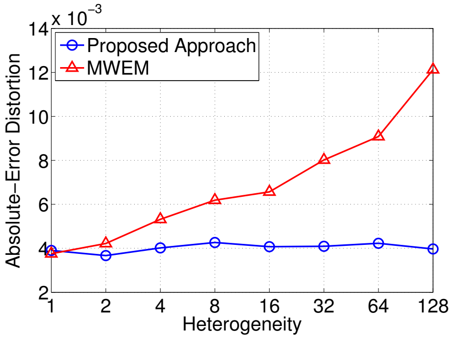

Fig. 2: Distortion under varying heterogeneity. The proposed approach is robust to heterogeneity, whereas the distortion of the MWEM algorithm grows as the heterogeneity increases.

statistical queries with heterogeneity h independently. We use the absolute-error distortion measure, i.e., $\rho(s,t) = |s-t|$ for any $s,t \in \mathbb{R}$, since both our approach and the MWEM algorithm have distortion upper bound under this distortion measure. We measure the worst-case distortion among the queries in the query set, and then take an average over 20 independent runs. The differential privacy level is fixed to $\epsilon = 1$.

Fig. 2 compares our approach against the MWEM algorithm with varying heterogeneity. The figure shows that the proposed approach gives similar distortions irrespective of the heterogeneity, whereas under the MWEM algorithm, the distortion grows as the heterogeneity increases. This experimental result shows a separation between statistical queries and linear queries: approaches designed for linear queries cannot be directly applied to statistical queries without performance loss.

2) Query Set Size–Independent Distortion: Under most existing mechanisms [4], [5], [9]–[11] for synthetic database release, the accuracy guarantee becomes worse as the query set size increases. Under the MWEM algorithm, the worst-case distortion among the queries in a query set is $O((\log(|\mathcal{Q}|))^{1/3})$, where $|\mathcal{Q}|$ is the query set size. Our approach does not restrict to a specific query set. The distortion upper bound holds for all the statistical queries. Therefore, under our approach, the worst-case distortion among the queries in a query set will not grow as the query set size increases.

We evaluate the proposed approach and the MWEM algorithm on databases from the Netflix dataset. We randomly generate linear query sets with the size varying from 64 to $1,048,576$, using the same method as in the previous experiments. We still use the absolute-error distortion measure. We measure the worst-case distortion among the queries in the query set and among 50 databases, with database sizes roughly within $1000 \sim 2000$. Then the worst-case distortion is averaged over 20 independent runs. The differential privacy level is fixed to $\epsilon = 1$.

Fig. 3 compares our approach against the MWEM algorithm with varying query set size. The figures shows that the proposed approach gives similar worst-case distortion for different query set sizes. However, for the MWEM algorithm, although very slowly, the worst-case distortion grows as the query set size increases. Therefore, to achieve certain accuracy, this growth indicates that the query set size must be smaller than a threshold. This experimental result verifies the dependence of the distortion on the query set size under the MWEM algorithm, and shows the advantage of our approach.

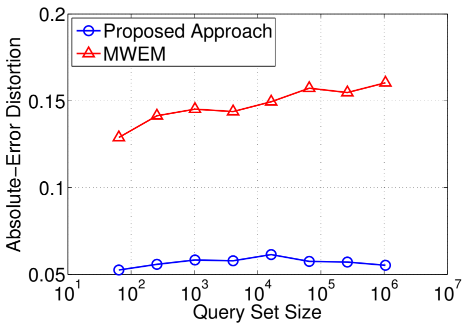

Fig. 3: Distortion under varying query set size. The worst-case distortion of the proposed approach does not depend on the query set size, whereas the distortion of the MWEM algorithm grows (slowly) as the query set size increases.

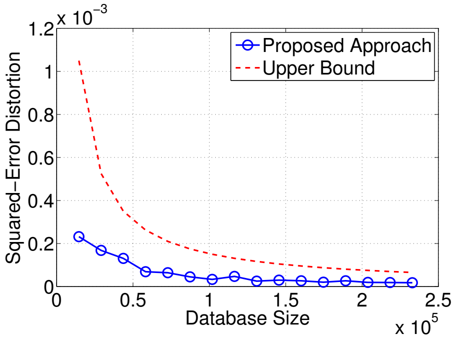

Fig. 4: Distortion under varying database size. In the asymptotic regime that the database size n goes to infinity, the upper bound is $\Theta(1/n)$, so the distortion is $O(1/n)$.

3) Scaling Behavior: Consider the asymptotic regime that the database size n goes to infinity for given data universe dimension and differential privacy level. We have proved that the worst-case squared-error distortion of the mechanism $\mathcal{E}$ when companioned with the estimator $\hat{q}^{\mathrm{u}}_{\varphi}$ is $O(1/n)$. To verify this theoretical upper bound, we evaluate the proposed approach on databases from the Netflix dataset. The sizes of the databases vary from $14,559$ to $232,944$. A linear query set of size 200 is randomly generated in the same way as the previous experiments and used for all the databases. We use the squared-error distortion measure, i.e., $\rho(s,t) = (s-t)^2$ for any $s,t \in \mathbb{R}$. We measure the worst-case distortion among the queries in the query set, and then take an average over 20 independent runs. The differential privacy level is fixed to $\epsilon = 1$. Fig. 4 compares the actual distortion under the proposed approach with the distortion upper bound, which verifies the asymptotic order $O(1/n)$ of the distortion.

B. *Differentially Private Cut Function Release for Graphs*

Consider the scenario that the given database is a graph, where the presence of individual edges is sensitive information. Such a graph can represent the online social connections between individuals. To release useful information for graph analysis, a well studied approach is to privately release the cut function of the graph [10], [24], [25].

Let the graph be $G = (V, E)$ and $\wp(V)$ denote the power set of V. Then the cut function $f_G \colon \wp(V) \times \wp(V) \to [|E|]$ associated with this graph is defined by

$$f_G(S, T) = |\{(i, j) \in E \mid i \in S, j \in T\}|, \tag{23}$$

which is the number of edges crossing the S, T-cut for any disjoint $S, T \subseteq V$.

We use a database x to represent the graph G. Since differential privacy needs to be preserved for edges, each row of x corresponds to a vertex pair $(i, j) \in V \times V$, where $x_{i,j} = 1$ if $(i, j) \in E$, and $x_{i,j} = 0$ otherwise. Here we use (i, j) to index each row of x. Thus the data universe is $\{0, 1\}$ with dimension $l = 1$ and the database size $n = |V|^2$. Two databases x, x' are neighbors if there exists exactly one vertex pair (i, j) such that $x_{i,j} \neq x'_{i,j}$.

For any disjoint $S, T \subseteq V$, we write $f_G(S, T)$ as a function $q_{S,T}$ of x and call it a *cut query*. Consider the absolute-error distortion measure $\rho(s, t) = |s - t|$ for any $s, t \in \mathbb{R}$. Then the minimax distortion for ϵ-differentially private cut function release can be written as

$$\mathfrak{D}_\epsilon^{\mathrm{C}} = \inf_{\mu_{\mathcal{M}} \in \mathcal{U}_\epsilon} \sup_{\substack{x \in \{0,1\}^n \\ S,T \subseteq V, S \cap T = \emptyset}} \mathbb{E}_{Y \sim \mu_{\mathcal{M}}(x)} \left[|\hat{q}^*_{S,T}(Y) - q_{S,T}(x)| \right].$$

Consider the statistical query defined in Definition 3. Then a cut query $q_{S,T}$ can be viewed as an unnormalized statistical query over the subset $S \times T \subseteq V \times V$ of all the rows. The row function is $\varphi_{i,j}(x_{i,j}) = x_{i,j}$ since

$$q_{S,T}(x) = \sum_{(i,j) \in S \times T} x_{i,j}. \tag{24}$$

Consider the mechanism $\mathcal{E}$ and estimator $\hat{q}_{S,T} \colon \{0, 1\}^n \to \mathbb{R}$ defined by

$$\hat{q}_{S,T}(y) = \frac{1 + e^{-\epsilon}}{1 - e^{-\epsilon}} q_{S,T}(y) - \frac{e^{-\epsilon}}{1 - e^{-\epsilon}} |S||T|, \tag{25}$$

which is an adapted version of the estimator $\hat{q}^{\mathrm{u}}_\varphi$ defined in (12) for the query $q_{S,T}$. Let Y denote the released synthetic database $\mathcal{E}(x)$. By similar analysis as in the proof of Lemma 2, the distortion is bounded as

$$\mathbb{E}_{Y \sim \mu_{\mathcal{E}}(x)} \left[|\hat{q}_{S,T}(Y) - q_{S,T}(x)| \right] \leq \frac{1 + e^{-\epsilon}}{1 - e^{-\epsilon}} \sqrt{|S||T|}. \tag{26}$$

For any $S, T \subseteq V$, $|S||T| \leq |V|^2$. Therefore the minimax distortion is upper bounded as

$$\mathfrak{D}_\epsilon^{\mathrm{C}} \leq \frac{1 + e^{-\epsilon}}{1 - e^{-\epsilon}} |V|. \tag{27}$$

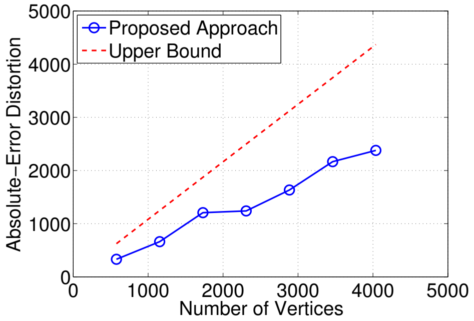

Fig. 5: Distortion of cut queries under varying number of nodes in the graph. In the asymptotic regime that the number of nodes $|V|$ goes to infinity, the upper bound is $\Theta(|V|)$, so the distortion is $O(|V|)$.

$\lvert V\rvert$	577	1154	1731	2308	2885	3462	4039
Error	10.4%	11.7%	8.7%	5.3%	4.7%	5.3%	5.4%

TABLE I: Relative error for cut queries.

1) Evaluation on the Facebook Dataset: We evaluate the proposed approach on databases from the Facebook dataset for the application of cut query release. The Facebook dataset is a graph. Each vertex in the graph represents a user, and an edge between two vertices indicates that they are friends.

Consider the asymptotic regime that the number of vertices $|V|$ goes to infinity. We have proved that the absolute-error distortion for any cut query is $O(|V|)$. To verify this theoretical upper bound, we apply our approach on subgraphs of the graph given by the Facebook dataset. The graph consists of 4039 vertices and $88,234$ edges. The number of vertices in the considered subgraphs vary from 577 to 4039. For each subgraph, cut queries are generated randomly in the following way. Half of the vertices are uniformly sampled and this vertex set is denoted by S. Then S and $V - S$ specify a cut query. This choice of cut queries results in the largest upper bound on the distortion as shown in (26). We generate a cut query set consisting of 100 cut queries independently. We measure the worst-case absolute-error distortion among the cut queries in the query set, and then take an average over 10 independent runs. The differential privacy level is fixed to $\epsilon = 1$. Fig. 5 compares the distortion under the proposed approach with the upper bound in (26), which verifies the asymptotic order $O(|V|)$ of the distortion. The worst-case relative distortion in Table I shows that the accuracy is reasonable for cut queries.

Appendix B
Proof of Lemma 2

Proof. We first prove the following claim.

Claim. *Under the mechanism $\mathcal{E}$, the estimator $\hat{q}^{\mathrm{u}}_\varphi$ is unbiased, i.e., for any database $x \in \mathcal{D}^n$,*

$$\mathbb{E}[\hat{q}^{\mathrm{u}}_\varphi(Y)] = q(x), \tag{28}$$

and the distortion of $\hat{q}_\varphi^{\mathrm{u}}$ satisfies the following upper bound:

$$\sup_{x\in\mathcal{D}^n} \mathbb{E}\big[|\hat{q}_\varphi^{\mathrm{u}}(Y) - q_\varphi(x)|^2\big] \le \frac{(b-a)^2\big(1+(2^l-1)e^{-\epsilon}\big)^2}{c^2(1-e^{-\epsilon})^2}\frac{1}{n}, \tag{29}$$

where a, b, c are the constants in Definition 3.

We drop the subscript $Y \sim \mu_{\mathcal{E}(x)}$ from expectations for conciseness during the proof. We first prove that the estimator $\hat{q}_\varphi^{\mathrm{u}}$ is unbiased. Recall that $\{Y_i, i\in[n]\}$ follow the PMFs in (11). Then

$$\begin{aligned}
\mathbb{E}[q_\varphi(Y)] &= \frac{1}{\sum_{i=1}^n c_i}\sum_{i=1}^n \mathbb{E}[\varphi_i(Y_i)] \\
&= \frac{1}{\sum_{i=1}^n c_i}\sum_{i=1}^n \bigg(\frac{1}{g(\epsilon)}\varphi_i(x_i) + \frac{e^{-\epsilon}}{g(\epsilon)}\sum_{\substack{v\in\mathcal{D}:\\ v\neq x_i}}\varphi_i(v)\bigg) \\
&= \frac{1-e^{-\epsilon}}{g(\epsilon)}\frac{1}{\sum_{i=1}^n c_i}\sum_{i=1}^n \varphi_i(x_i) \\
&\quad + \frac{e^{-\epsilon}}{g(\epsilon)}\frac{1}{\sum_{i=1}^n c_i}\sum_{i=1}^n\sum_{v\in\mathcal{D}}\varphi_i(v) \\
&= \frac{1-e^{-\epsilon}}{g(\epsilon)}q_\varphi(x) + \frac{e^{-\epsilon}}{g(\epsilon)}C_\varphi.
\end{aligned}$$

Therefore

$$\mathbb{E}[\hat{q}_\varphi^{\mathrm{u}}(Y)] = \mathbb{E}\bigg[\frac{g(\epsilon)}{1-e^{-\epsilon}}q_\varphi(Y) - \frac{e^{-\epsilon}}{1-e^{-\epsilon}}C_\varphi\bigg] = q_\varphi(x).$$

Next we prove the upper bound on the distortion of $\hat{q}_\varphi^{\mathrm{u}}$. For any $x\in\mathcal{D}^n$,

$$\begin{aligned}
&\hat{q}_\varphi^{\mathrm{u}}(Y) - q_\varphi(x) \\
&= \frac{g(\epsilon)}{1-e^{-\epsilon}}\frac{1}{\sum_{i=1}^n c_i} \\
&\quad\cdot \sum_{i=1}^n\bigg(\varphi_i(Y_i) - \frac{1-e^{-\epsilon}}{g(\epsilon)}\varphi_i(x_i) - \frac{e^{-\epsilon}}{g(\epsilon)}\sum_{v\in\mathcal{D}}\varphi_i(v)\bigg).
\end{aligned}$$

For any $i\in[n]$, let

$$Z_i = \varphi_i(Y_i) - \frac{1-e^{-\epsilon}}{g(\epsilon)}\varphi_i(x_i) - \frac{e^{-\epsilon}}{g(\epsilon)}\sum_{v\in\mathcal{D}}\varphi_i(v).$$

Then for any $i\in[n]$, $\mathbb{E}[Z_i]=0$. Recall that for any $v\in\mathcal{D}$, $a\le\varphi_i(v)\le b$, so $|Z_i|\le b-a$. Since $Y_1,\dots,Y_n$ are independent, $Z_1,\dots,Z_n$ are independent. Let $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$. Then

$$\begin{aligned}
&\mathbb{E}\big[|\hat{q}_\varphi^{\mathrm{u}}(Y) - q_\varphi(x)|^2\big] \\
&= \bigg(\frac{g(\epsilon)}{1-e^{-\epsilon}}\frac{n}{\sum_{i=1}^n c_i}\bigg)^2\cdot\mathbb{E}\big[\big|\overline{Z}\big|^2\big] \\
&= \bigg(\frac{g(\epsilon)}{1-e^{-\epsilon}}\frac{n}{\sum_{i=1}^n c_i}\bigg)^2\cdot\bigg(\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\big[|Z_i|^2\big]\bigg) \\
&\le \bigg(\frac{g(\epsilon)}{1-e^{-\epsilon}}\bigg)^2\frac{1}{c^2}\frac{(b-a)^2}{n}.
\end{aligned}$$

Therefore

$$\sup_{x\in\mathcal{D}^n} \mathbb{E}\big[|\hat{q}_\varphi^{\mathrm{u}}(Y) - q_\varphi(x)|^2\big] \le \frac{(b-a)^2\big(1+(2^l-1)e^{-\epsilon}\big)^2}{c^2(1-e^{-\epsilon})^2}\frac{1}{n},$$

which completes the proof of the claim.

Next we quantify the distortion of the proper estimator q_φ. For any $x,y\in\mathcal{D}^n$, since $q_\varphi(x)\in q_\varphi(\mathcal{D}^n)$, by the definition of the estimator $\hat{q}_\varphi$ in (14),

$$|\hat{q}_\varphi^{\mathrm{u}}(y) - \hat{q}_\varphi(y)| \le |\hat{q}_\varphi^{\mathrm{u}}(y) - q_\varphi(x)|.$$

Therefore

$$\begin{aligned}
|\hat{q}_\varphi(y) - q_\varphi(x)| &\le |\hat{q}_\varphi(y) - \hat{q}_\varphi^{\mathrm{u}}(y)| + |\hat{q}_\varphi^{\mathrm{u}}(y) - q_\varphi(x)| \\
&\le 2|\hat{q}_\varphi^{\mathrm{u}}(y) - q_\varphi(x)|,
\end{aligned}$$

and

$$\mathbb{E}\big[|\hat{q}_\varphi(Y) - q_\varphi(x)|^2\big] \le 4\mathbb{E}\big[|\hat{q}_\varphi^{\mathrm{u}}(Y) - q_\varphi(x)|^2\big].$$

Then combining with (29) yields the upper bound. □

Appendix C
Example Illustrating How to Obtain C_φ

By the form of the estimator $\hat{q}_\varphi^{\mathrm{u}}$ in (12), the value

$$C_\varphi = \frac{1}{\sum_{i=1}^n c_i}\sum_{i=1}^n\sum_{v\in\mathcal{D}}\varphi_i(v)$$

is needed to answer the query q_φ. However, in many cases, this value can be easily obtained rather than exhaustive calculation. In such case, the computation in $\hat{q}_\varphi^{\mathrm{u}}$ is very efficient. Take the following predicate query for an example. Recall that any $v\in\mathcal{D}=\{0,1\}^l$ is a binary vector $v=(v_1,\dots,v_l)$ of length l. Consider the predicate function $s(v) = v_{j_1}\cdot v_{j_2}\cdot\ldots v_{j_k}$ for some $\{j_1,\dots,j_k\}$ with $1\le k\le l$, which counts the fraction of rows in the database that have value 1 for attributes $j_1,\dots,j_k$. This predicate query is a statistical query q_φ with $\varphi_i = s$ for any $i\in[n]$. The value C_φ for this query is $C_\varphi = 2^{l-k}$, which can be obtained by simple analysis.

Appendix D
Proof of Lemma 3

Proof. By Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z] - d(X,Z)|^2\big] \\
&\ge \big(\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z] - d(X,Z)|\big]\big)^2.
\end{aligned} \tag{30}$$

Let $\widetilde{X}$ be a random variable satisfying the following conditions: $\widetilde{X}$ is independent of Z; $\widetilde{X}$ is independent of X given Y; given Y, $\widetilde{X}$ and X are identically distributed, i.e., $p_{\widetilde{X}|Y}(x\mid y) = p_{X|Y}(x\mid y)$ for any $x,y\in\mathcal{D}^n$ with $p_Y(y)\neq 0$. Due to the independence between Z and $(X,Y,\widetilde{X})$, we also have $p_{\widetilde{X}|Y,Z}(x\mid y,z) = p_{X|Y,Z}(x\mid y,z)$ for any $x,y,z\in\mathcal{D}^n$ with $p_Y(y)\neq 0$. By this construction, for any $y,z\in\mathcal{D}^n$ with $p_Y(y)\neq 0$,

$$\mathbb{E}[d(X,Z)\mid Y=y,Z=z] = \mathbb{E}[d(\widetilde{X},Z)\mid Y=y,Z=z],$$

and

$$\begin{aligned}&\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z]-d(X,Z)|\;\big|\;Y=y,Z=z\big]\\ &=\mathbb{E}\big[|\mathbb{E}[d(\widetilde{X},Z)\mid Y,Z]-d(\widetilde{X},Z)|\;\big|\;Y=y,Z=z\big],\end{aligned}$$

which further lead to

$$\begin{aligned}&\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z]-d(X,Z)|\big]\\ &=\mathbb{E}\Big[\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z]-d(X,Z)|\;\big|\;Y,Z\big]\Big]\\ &=\mathbb{E}\Big[\mathbb{E}\big[|\mathbb{E}[d(\widetilde{X},Z)\mid Y,Z]-d(\widetilde{X},Z)|\;\big|\;Y,Z\big]\Big]\\ &=\mathbb{E}\big[|\mathbb{E}[d(\widetilde{X},Z)\mid Y,Z]-d(\widetilde{X},Z)|\big].\end{aligned}$$

Therefore

$$\begin{aligned}&2\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z]-d(X,Z)|\big]\\ &=\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z]-d(X,Z)|\\ &\qquad+|\mathbb{E}[d(\widetilde{X},Z)\mid Y,Z]-d(\widetilde{X},Z)|\big]\\ &\geq\mathbb{E}\big[|d(X,Z)-d(\widetilde{X},Z)\\ &\qquad+\mathbb{E}[d(\widetilde{X},Z)\mid Y,Z]-\mathbb{E}[d(X,Z)\mid Y,Z]|\big]\\ &=\mathbb{E}\big[|d(X,Z)-d(\widetilde{X},Z)|\big].\end{aligned}$$

Combing this with (30) gives

$$\begin{aligned}&\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z]-d(X,Z)|^2\big]\\ &\geq\frac{1}{4}\big(\mathbb{E}\big[|d(X,Z)-d(\widetilde{X},Z)|\big]\big)^2.\end{aligned}\tag{31}$$

Then it suffices to derive a lower bound on $\mathbb{E}\big[|d(X,Z)-d(\widetilde{X},Z)|\big]$.

Notice that the conditional PMF $p_{\widetilde{X}|X}$ is ϵ-differentially private since for any neighboring $x,x'\in\mathcal{D}^n$ and any $\widetilde{x}\in\mathcal{D}^n$,

$$\begin{aligned}p_{\widetilde{X}|X}(\widetilde{x}\mid x)&=\sum_{y\in\mathcal{D}^n}p_{\widetilde{X}|Y,X}(\widetilde{x}\mid y,x)p_{Y|X}(y\mid x) &(32)\\ &=\sum_{y\in\mathcal{D}^n}p_{\widetilde{X}|Y,X}(\widetilde{x}\mid y,x')p_{Y|X}(y\mid x) &(33)\\ &\leq\sum_{y\in\mathcal{D}^n}p_{\widetilde{X}|Y,X}(\widetilde{x}\mid y,x')\cdot e^{\epsilon}p_{Y|X}(y\mid x') &(34)\\ &=e^{\epsilon}p_{\widetilde{X}|X}(\widetilde{x}\mid x'), &(35)\end{aligned}$$

where (33) follows from the conditional independence between $\widetilde{X}$ and X given Y, and (34) holds because $p_{Y|X}$ is ϵ-differentially private. Then by Theorem 1 in [12] (for our case, the ϵ_X in that theorem is 0),

$$\mathbb{E}[d(X,\widetilde{X})]\geq\frac{n}{1+\frac{e^{\epsilon}}{2^l-1}}.$$

Let $\gamma=\frac{1}{2(1+\frac{e^{\epsilon}}{2^l-1})}$ and $s=\gamma n$. Since

$$\begin{aligned}\mathbb{E}[d(X,\widetilde{X})]&\leq s\mathbb{P}\{d(X,\widetilde{X})<s\}+n\mathbb{P}\{d(X,\widetilde{X})\geq s\}\\ &\leq s+n\mathbb{P}\{d(X,\widetilde{X})\geq s\},\end{aligned}$$

we have

$$\begin{aligned}\mathbb{P}\{d(X,\widetilde{X})\geq s\}&\geq\frac{1}{n}(\mathbb{E}[d(X,\widetilde{X})]-s)\\ &\geq\frac{1}{n}\left(\frac{n}{1+\frac{e^{\epsilon}}{2^l-1}}-\frac{n}{2\big(1+\frac{e^{\epsilon}}{2^l-1}\big)}\right)\\ &=\gamma,\end{aligned}$$

i.e.,

$$\mathbb{P}\{d(X,\widetilde{X})\geq\gamma n\}\geq\gamma.\tag{36}$$

We will consider those $x,\widetilde{x}\in\mathcal{D}^n$ with $d(x,\widetilde{x})\geq\gamma n$ to obtain a lower bound on $\mathbb{E}[|d(X,Z)-d(\widetilde{X},Z)|]$.

Utilizing conditional expectation gives

$$\begin{aligned}&\mathbb{E}[|d(X,Z)-d(\widetilde{X},Z)|]\\ &=\mathbb{E}\big[\mathbb{E}[|d(X,Z)-d(\widetilde{X},Z)|\mid X,\widetilde{X}]\big]\\ &\geq\sum_{\substack{x,\widetilde{x}:\\ d(x,\widetilde{x})\geq\gamma n}}\mathbb{E}[|d(X,Z)-d(\widetilde{X},Z)|\mid X=x,\widetilde{X}=\widetilde{x}]p_{X,\widetilde{X}}(x,\widetilde{x}).\end{aligned}\tag{37}$$

Consider any $x,\widetilde{x}\in\mathcal{D}^n$ with $d(x,\widetilde{x})\geq\gamma n$ and $p_{X,\widetilde{X}}(x,\widetilde{x})\neq 0$. Since Z is independent of $(X,\widetilde{X})$,

$$\begin{aligned}&\mathbb{E}[|d(X,Z)-d(\widetilde{X},Z)|\mid X=x,\widetilde{X}=\widetilde{x}]\\ &=\mathbb{E}[|d(x,Z)-d(\widetilde{x},Z)|].\end{aligned}\tag{38}$$

Let

$$\Delta(x,\widetilde{x})=\{i\in[n]\mid x_i\neq\widetilde{x}_i\}.\tag{39}$$

Then $|\Delta(x,\widetilde{x})|\geq\gamma n$, and

$$\begin{aligned}|d(x,Z)-d(\widetilde{x},Z)|&=\left|\sum_{i=1}^{n}\big(\zeta(x_i,Z_i)-\zeta(\widetilde{x}_i,Z_i)\big)\right|\\ &=\left|\sum_{i\in\Delta(x,\widetilde{x})}\big(\zeta(x_i,Z_i)-\zeta(\widetilde{x}_i,Z_i)\big)\right|.\end{aligned}$$

Let

$$U_i=\zeta(x_i,Z_i)-\zeta(\widetilde{x}_i,Z_i).\tag{40}$$

Since Z is uniformly distributed over $\mathcal{D}^n$, the rows $Z_1,Z_2,\ldots,Z_n$ are i.i.d. with PMF $p_{Z_i}(z_i)=\frac{1}{2^l}$ for any $z_i\in\mathcal{D}$. For any $i\in\Delta(x,\widetilde{x})$,

$$U_i=\begin{cases}1 & \text{if } Z_i=\widetilde{x}_i,\\ -1 & \text{if } Z_i=x_i,\\ 0 & \text{otherwise.}\end{cases}\tag{41}$$

Therefore $\{U_i,i\in\Delta(x,\widetilde{x})\}$ are i.i.d. with PMF

$$p_{U_i}(u_i)=\begin{cases}\frac{1}{2^l} & u_i=1,\\ \frac{1}{2^l} & u_i=-1,\\ 1-\frac{1}{2^{l-1}} & u_i=0.\end{cases}\tag{42}$$

Then $\mathbb{E}[U_i]=0$. Denote

$$\sigma^2=\mathbb{E}\big[|U_i|^2\big]=\frac{1}{2^{l-1}},\quad\rho=\mathbb{E}\big[|U_i|^3\big]=\frac{1}{2^{l-1}}.\tag{43}$$

By the Berry–Esseen theorem [26, Theorem 7.4.1], there exists a universal constant C such that for any t,

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{\sigma\sqrt{|\Delta(x,\widetilde{x})|}}\sum_{i\in\Delta(x,\widetilde{x})} U_i > \frac{t}{\sigma\sqrt{\gamma}}\right\} \\
&\geq 1-\Phi\Big(\frac{t}{\sigma\sqrt{\gamma}}\Big)-\frac{C\rho}{\sigma^3\sqrt{|\Delta(x,\widetilde{x})|}} \\
&\geq 1-\Phi\Big(\frac{t}{\sigma\sqrt{\gamma}}\Big)-\frac{C\rho}{\sigma^3\sqrt{\gamma n}},
\end{aligned}$$

where the second inequality follows from $|\Delta(x,\widetilde{x})| \geq \gamma n$. Therefore

$$\begin{aligned}
&\mathbb{P}\big\{|d(x,Z)-d(\widetilde{x},Z)| > t\sqrt{n}\big\} \\
&= \mathbb{P}\left\{\frac{1}{\sigma\sqrt{|\Delta(x,\widetilde{x})|}}\sum_{i\in\Delta(x,\widetilde{x})} U_i > \frac{t\sqrt{n}}{\sigma\sqrt{|\Delta(x,\widetilde{x})|}}\right\} \\
&\geq \mathbb{P}\left\{\frac{1}{\sigma\sqrt{|\Delta(x,\widetilde{x})|}}\sum_{i\in\Delta(x,\widetilde{x})} U_i > \frac{t\sqrt{n}}{\sigma\sqrt{\gamma n}}\right\} \\
&\geq 1-\Phi\Big(\frac{t}{\sigma\sqrt{\gamma}}\Big)-\frac{C\rho}{\sigma^3\sqrt{\gamma n}}.
\end{aligned}$$

Let $t=\sigma\sqrt{\gamma}$, then

$$\mathbb{P}\big\{|d(x,Z)-d(\widetilde{x},Z)| > \sigma\sqrt{\gamma n}\big\} \geq 1-\Phi(1)-\frac{C\rho}{\sigma^3\sqrt{\gamma n}},$$

and further

$$\begin{aligned}
&\mathbb{E}[|d(x,Z)-d(\widetilde{x},Z)|] \\
&\geq \sigma\sqrt{\gamma n}\cdot\mathbb{P}\big\{|d(x,Z)-d(\widetilde{x},Z)| > \sigma\sqrt{\gamma n}\big\} \\
&\geq \big(1-\Phi(1)\big)\sigma\sqrt{\gamma n}-\frac{C\rho}{\sigma^3}. \qquad (44)
\end{aligned}$$

Inserting this lower bound back to (38), (37) and combining the lower bound (36) yield

$$\begin{aligned}
&\mathbb{E}[|d(X,Z)-d(\widetilde{X},Z)|] \\
&\geq \sum_{\substack{x,\widetilde{x}:\\ d(x,\widetilde{x})\geq\gamma n}}\left(\big(1-\Phi(1)\big)\sigma\sqrt{\gamma n}-\frac{C\rho}{\sigma^3}\right)p_{X,\widetilde{X}}(x,\widetilde{x}) \\
&= \left(\big(1-\Phi(1)\big)\sigma\sqrt{\gamma n}-\frac{C\rho}{\sigma^3}\right)\mathbb{P}\{d(X,\widetilde{X})\geq\gamma n\} \\
&\geq \big(1-\Phi(1)\big)\sigma\gamma^{\frac{3}{2}}\sqrt{n}-\frac{C\rho\gamma}{\sigma^3}.
\end{aligned}$$

Therefore, by (31),

$$\begin{aligned}
&\mathbb{E}\big[|\mathbb{E}[d(X,Z)\mid Y,Z]-d(X,Z)|^2\big] \\
&\geq \frac{1}{4}\left(\big(1-\Phi(1)\big)\sigma\gamma^{\frac{3}{2}}\sqrt{n}-\frac{C\rho\gamma}{\sigma^3}\right)^2,
\end{aligned}$$

which completes the proof. □