Partial Data Extraction via Noisy Histogram Queries: Information Theoretic Bounds

Wei-Ning Chen and I-Hsiang Wang
Graduate Institute of Communication Engineering and Department of Electrical Engineering,
National Taiwan University, Taipei, Taiwan
Email: {r05942078,ihwang}@ntu.edu.tw

Abstract*—The problem of extracting categorical data via noisy histogram queries is investigated. The considered data set is a collection of n items, each of which carries a piece of categorical data taking values in a finite alphabet. Data analysts are allowed to query the data set through a curator by specifying a subset of items and then obtaining the histogram of the queried subset. The (unnormalized) histogram released by the curator, however, is perturbed by some additive noise with maximum magnitude δ_n. The goal of the data analyst is to reconstruct the categorical data set such that the Hamming distance between the reconstructed and the actual one is smaller than a tolerance parameter k_n. In this work, we explore the fundamental limit on the minimum number of queries T_n^* required for the analyst to reconstruct the n-item data set within k_n tolerance subject to δ_n noisy perturbation. We first show that if $\delta_n = O(\sqrt{k_n})$, the minimum query complexity $T_n^* = \Theta(n/\log n)$, where the achievability is based on random sampling, and the converse is based on counting and packing arguments. On the other hand, if $\delta_n = \Omega(k_n^{(1+\epsilon)/2})$ for some $\epsilon > 0$, we prove that $T_n^* = \omega(n^p)$ for any positive integer p. In other words, no querying methods with polynomial-in-n query complexity can successfully reconstruct the data set in that regime. This impossibility result is established by a novel combinatorial lower bound on T_n^*.

I. Introduction

Extracting information from large-scale data sets plays a crucial role in many fields including data mining, machine learning, bioinformatics, etc.. Typically, data extraction consists of two steps: first the data analyst sends *queries* to the curator who is in charge of data release, and then the data curator responds with the corresponding answers. In many circumstances the data curator only allows certain types of queries whose outcomes do not depend on individual items, due to privacy considerations or computational efficiency limitations. For categorical data, a typical kind of query is *histogram query*, where each query is a subset of items, and the response is the histogram (the number of items belonging to each category) of the corresponding items. In addition to constraining the type of queries, the curator may further perturb the response to provide stronger privacy guarantees. For example, in [1], histogram query (or called *counting query* for binary category) is studied and analyzed as a privacy-preserving database model. Therefore, in this paper we focus on noisy responses to histogram queries.

Characterizing the fundamental limit on the number of queries (termed *query complexity*) required to extract the data set is important to both data analysts and data curators. In [2], the fundamental limit on the minimum query complexity to precisely extract the entire n-item data set with noiseless histogram queries is characterized. The optimal query complexity was shown to be $\Theta(n/\log n)$, where n is the size of the data set. Moreover, an explicit construction of the querying method achieving the optimal query complexity is proposed. However, for the general setting where the goal is to *partially* extract the data set with *noisy* query responses, the characterization of the optimal query complexity remains open.

In this paper, we investigate the optimal query complexity T_n^* for partial data extraction with noisy responses to histogram queries. The response from the curator is the actual unnormalized histogram of the queried subset of items, perturbed by an additive noise with maximum magnitude δ_n. The goal of the analyst is to reconstruct the data set partially so that the Hamming distance between the reconstructed and the actual data set is at most k_n. Our main contribution is characterizing the asymptotic behavior of T_n^* with respect to the size of the data set n and the two parameters k_n, δ_n coupled with n:

1) In the regime $\delta_n = O(\sqrt{k_n})$, $T_n^* = \Theta(n/\log n)$, which is the same as the optimal query complexity for perfect reconstruction with noiseless responses to queries [2].
2) In the regime $\delta_n = \Omega(k_n^{(1+\epsilon)/2})$ for some $\epsilon > 0$, $T_n^* = \omega(n^p)$ for any positive integer p. In words, there does not exist querying methods with $\mathsf{Poly}(n)$ query complexity.

For proving the achievability part (upper bound on T_n^*), *randomized* querying is employed. In each query, the items to be included in the queried subset are randomly and uniformly selected. An upper bound on the probability of failure to distinguish two different data sets is then proved, showing that if $\delta_n = O(\sqrt{k_n})$, $\Omega(n/\log n)$ such queries ensure vanishing probability of failure. For proving the converse part (lower bound on T_n^*), we first show that $T_n^* = \Omega(n/\log n)$ based on a packing argument, extending the proof in [2] to general δ_n, k_n. We then develop a novel combinatorial lower bound on T_n^* and show that if $\delta_n = \Omega(k_n^{(1+\epsilon)/2})$ for some $\epsilon > 0$ then no method with polynomial query complexity can reconstruct the data set within Hamming distance of k_n.

Related Works: Prior work on categorical data extraction with histogram queries for generic alphabet $\mathcal{A}$ was initiated in [2], where the optimal query complexity of exact reconstruction is shown to be $\Theta(n/\log n)$ with noiseless responses, and improved to $\Theta(\frac{k}{\log k}\log\frac{n}{k})$ when the data set is sparse with sparsity level k [3]. Furthermore in [4], upper and lower bounds on the pre-constants in the $n/\log n$ scaling are also proved. However, none of the previous works investigated the setting with noisy responses to queries and partial reconstruction. Our problem can also be viewed as generalization of *group testing*. See Section VI. of [2] for the connection.

Our work is closely related to studies of lower bounds in *data privacy*, where the focus is on deriving conditions on the perturbation level in the response so that no *computationally-efficient* algorithms can reconstruct the private data set from aggregated queries. Binary alphabet ($\mathcal{A} = \{0,1\}$) is mainly considered in these works. In [1], noisy response to histogram query is proven to be *differential private* with proper perturbation. In [5], it is shown that no algorithm with polynomial running time can reconstruct a constant fraction of the entire data set when the perturbation level $\delta_n = \Omega(\sqrt{n})$. Besides, when $\delta_n = o(\sqrt{n})$, a polynomial-running-time algorithm is given, where the query complexity is $\omega(n)$. In [6], query complexity and running time are improved to n and $\Theta(n \log n)$ respectively. However, all the reconstruction algorithms [5]–[8] aim to recover only a constant fraction of the entire data set ($k_n = \Theta(n)$) with perturbation $\delta_n \approx \sqrt{n}$, and can be viewed as special cases in the regimes considered in this work.

Notations: $[N_1 : N_2] \triangleq \{N_1, N_1+1, ..., N_2\}$ for integers $N_1 \le N_2$, and $[N] \triangleq \{1, 2, ..., N\}$, for $N \in \mathbb{N}$. Let $(\cdot)^\intercal$ denote the matrix transpose and $\mathbb{1}\{\cdot\}$ denote the indicator function.

II. Problem Formulation

Following [2], we cast the data extraction problem with n items and T_n queries as a linear inverse problem.

A. Data Set, Queries, and Responses

Consider a data set with n items, labeled from 1 to n. Each item possesses a piece of data which takes value in a finite alphabet $\mathcal{A} = \{a_1, a_2, ..., a_d\}$ and $|\mathcal{A}| = d$. We first consider the case $d = 2$, and assume without loss of generality $\mathcal{A} = \{0, 1\}$. Later in Section VI, it is explained how to extend the results to general d. Let us denote the data set as $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X}$ denotes the collection of all possible realization of data sets. For now, $\mathcal{X} = \{0,1\}^{n \times 1}$.

To address the partial reconstruction criterion, we use the Hamming distance, formally stated below.

Definition 2.1 (Distance between two data sets): Let $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{X}$ be two data sets with items $[x_1 \; ... \; x_n]^\intercal$ and $[\tilde{x}_1 \; ... \; \tilde{x}_n]^\intercal$ respectively. Then, $d_{\text{data}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \triangleq \sum_{j=1}^{n} \mathbb{1}\{x_j \neq \tilde{x}_j\}$.

Consider T_n queries, each query being a subset of labels in $[n]$. Let $\mathcal{S}_i$ denote the queried subset in the i-th query. The response to a query is the histogram of the queried subset. We shall use a $T_n \times n$ query matrix $\mathbf{Q} \in \{0,1\}^{T_n \times n}$ to collectively represent the T_n queries. In particular, $(\mathbf{Q})_{i,j} = 1$ if and only if the j-th item is included in the i-th queried subset. In other words, $(\mathbf{Q})_{i,j} = \mathbb{1}\{j \in \mathcal{S}_i\}$. Hence, the i-th row $\boldsymbol{q}_i^\intercal \in \{0,1\}^{1 \times n}$ represents the queried subset in the i-th query. The responses to the queries can then be represented as the multiplication of the query matrix and the data-set matrix (here, it is an $n \times 1$ matrix). It is not hard to see that the *unnormallized* response to the i-th histogram query $y_i = \boldsymbol{q}_i^\intercal \boldsymbol{x} \in [n]$ and hence $\boldsymbol{y} = \mathbf{Q}\boldsymbol{x} \in \{0, 1, ..., n\}^{T_n \times 1}$.

To address the perturbation in the responses, we use the ℓ_∞ norm, formally stated below.

Definition 2.2 (Distance between two response): Suppose $\boldsymbol{y}, \tilde{\boldsymbol{y}}$ are the responses to two queries. The distance between them is defined as $d_{\text{response}}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) \triangleq \max_i |y_i - \tilde{y}_i|$.

B. Criteria of Data Extraction

Definition 2.3 (Tolerance in Partial Extraction): The data extraction task is *k-tolerable*, if the reconstructed data set $\tilde{\boldsymbol{x}}$ differs from the original one $\boldsymbol{x}$ by at most k, that is,

$$d_{\text{data}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \le k, \; \forall \boldsymbol{x} \in \mathcal{X}.$$

Definition 2.4 (Noise Level in Perturbed Response): Responses to queries is of noise level δ if the perturbed response $\tilde{\boldsymbol{y}}$ has distance at most δ to original $\boldsymbol{y}$, that is,

$$d_{\text{response}}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) \le \delta.$$

The goal of the data analyst is to design the query matrix $\mathbf{Q}$ to extract the data set $\boldsymbol{x}$ within distance k_n from the δ_n perturbed response $\tilde{\boldsymbol{y}}$. Formally, $\mathbf{Q}$ has to satisfy the following:

$$\begin{aligned} &\forall \boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{X}, \; d_{\text{data}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) > k_n \\ &\qquad \Longrightarrow d_{\text{response}}(\mathbf{Q}\boldsymbol{x}, \mathbf{Q}\tilde{\boldsymbol{x}}) > 2\delta_n. \end{aligned} \tag{1}$$

Definition 2.5 (Recoverability): Suppose a query matrix $\mathbf{Q} \in \{0,1\}^{T_n \times n}$ satisfies (1) with respect to tolerance k_n and noise level δ_n, then it is called (T_n, k_n, δ_n)-recoverable.

Definition 2.6 (Optimal Query Complexity): $T_n^*(k_n, \delta_n)$ denotes the minimum query complexity for reconstructing a n-item data set with tolerance k_n under noise level δ_n, that is,

- There exists a $\mathbf{Q}$ which is (T_n^*, k_n, δ_n)-recoverable.
- For all $T_n < T_n^*$, there does not exist query matrix $\mathbf{Q}$ which is (T_n, k_n, δ_n)-recoverable.

For a randomized querying method, the query matrix $\mathbf{Q}$ is randomly selected from distribution $P_{\mathbf{Q}}$. To specify the criterion of successfully extracting the data set under randomized querying, let us define the *probability of failure* as follows:

Definition 2.7 (Probability of Failure): For a data set $\boldsymbol{x} \in \mathcal{X}$, the probability of failure $P_f(\boldsymbol{x}; k_n, \delta_n)$ with respect to the randomly generated query matrix $\mathbf{Q}$ is defined as

$$P_{\mathbf{Q}}\left\{\exists\, \tilde{\boldsymbol{x}}, \; d_{\text{data}}(\tilde{\boldsymbol{x}}, \boldsymbol{x}) > k_n, d_{\text{response}}(\mathbf{Q}\tilde{\boldsymbol{x}}, \; \mathbf{Q}\boldsymbol{x}) \le 2\delta_n\right\}$$

Definition 2.8 ((T_n, k_n, δ_n)-achievable): Given a sequence of randomly generated query matrices $\{\mathbf{Q}^{(T_n, n)} \mid n \in \mathbb{N}\}$, we say it is (T_n, k_n, δ_n)-achievable, if

$$\lim_{n \to \infty} \max_{\boldsymbol{x} \in \mathcal{X}} P_f(\boldsymbol{x}; k_n, \delta_n) = 0 \tag{2}$$

III. Main Results

A. Achievability

Theorem 3.1: (Achievability of Randomized Querying) Suppose one generates the query matrix $\mathbf{Q}_{i,j}^{(T_n, n)}$ according to the following distribution:

$$\left(\mathbf{Q}^{(T_n, n)}\right)_{i,j} \overset{\text{i.i.d.}}{\sim} \operatorname{Ber}\left(\tfrac{1}{2}\right). \tag{3}$$

Then, the extraction criterion (2) will be satisfied as long as $T_n = \Omega(\frac{n}{\log n})$ and one of the following conditions holds:

1) $k_n = O(n^{\epsilon})$, for some $\epsilon < 1$ and $\delta_n = O(\sqrt{k_n})$

2) $\forall\, \epsilon < 1,\ k_n = \omega(n^{\epsilon})$ and $\delta_n = O(k_n^{\frac{1-\epsilon'}{2}})$ for some $\epsilon' > 0$.

Proof. The proof involves finding upper bounds on the probability of failure. Details can be found in Section IV. ■

B. Lower Bounds on Query Complexity

For the converse part, we give two lower bounds in the following two theorems.

Theorem 3.2: (Packing Lower Bound) Let $k_n \leq \left(\frac{1-\epsilon}{2}\right) n$ for some $\epsilon > 0$. Then, the following lower bound holds:

$$T_n^*(k_n, \delta_n) = \Omega\left(\frac{n\left(1 - H_b\left(\frac{1-\epsilon}{2}\right)\right)}{\log(n+1) - \log(4\delta_n + 1)}\right) \tag{4}$$

Specifically, when $\delta_n = O(n^{\frac{1-\epsilon'}{2}})$, and ϵ, ϵ' does not depend on n, then (4) can be further simplified to

$$T_n^*(k_n, \delta_n) = \Omega\left(n/\log n\right).$$

Proof. The successful extraction criterion holds only if for any two data sets $\boldsymbol{x}, \tilde{\boldsymbol{x}}$ with distance greater than k_n, the queried output $\mathbf{Q}\boldsymbol{x}, \mathbf{Q}\tilde{\boldsymbol{x}}$ differ to each other more than $2\delta_n$, say, $d_{\text{response}}(\mathbf{Q}\boldsymbol{x}, \mathbf{Q}\tilde{\boldsymbol{x}}) > 2\delta_n$. Therefore, we cast the problem into a packing problem. The detailed proof is omitted here and can be found in Appendix A of [9] ■

Remark 3.1: The condition $k_n \leq \left(\frac{1-\epsilon}{2}\right) n$ for some $\epsilon > 0$ is reasonable. Let $k_n = n/2$ and consider the following scenario: we simply make a query with $\boldsymbol{q}_i = [1, ..., 1]^{\mathsf{T}}$, and if $\boldsymbol{q}_i^{\mathsf{T}}\boldsymbol{x} > n/2$, we reconstruct $\boldsymbol{x}$ as $\tilde{\boldsymbol{x}} = [1, ..., 1]^{\mathsf{T}}$, else we say $\tilde{\boldsymbol{x}} = [0, ..., 0]^{\mathsf{T}}$. The reconstruction will succeed with high probability as n grows large enough, by making *one* query.

The above lower bound is used when the noise level δ_n is relatively small with respect to k_n. Next, we give another lower bound which depends on both k_n and δ_n:

Theorem 3.3: (Combinatorial Lower Bound)

$$T_n^*(k_n, \delta_n) \geq \frac{\binom{n}{n/2}}{2 \sum\limits_{\alpha=2\delta_n}^{k_n/2} \binom{k_n/2}{\alpha} \sum\limits_{\delta=2\delta_n}^{\alpha} \binom{k_n/2}{\alpha-\delta}\binom{n-k_n}{n/2-2\alpha+\delta}}. \tag{5}$$

This bound is used to prove the impossibility result when δ_n is large with respect to k_n. Detailed proof is given in Section V.

C. Fundamental Limit

First, Theorem 3.1 gives us a sufficient condition for recovering the data set by $\Omega(n/\log n)$ queries. On the other hand, Theorem 3.2 states that $T_n = \Omega(n/\log n)$ is also necessary for reconstruction. We combine them into the following corollary:

Corollary 3.1: (Fundamental Limit of Query Complexity) Under the one of the following two noise-tolerance conditions

- $k_n = O\left(n^{\epsilon}\right)$ for some $\epsilon < 1$, and $\delta_n = O(\sqrt{k_n})$, or
- $\forall\, \epsilon < 1,\ k_n = \omega\left(n^{\epsilon}\right)$, and $\exists\, \epsilon' > 0,\ \delta_n = O(k_n^{(1-\epsilon')/2})$,

the optimal query complexity is

$$T_n^*(k_n, \delta_n) = \Theta(\tfrac{n}{\log n}).$$

Next, following Theorem 3.3, we give an impossibility result below:

Theorem 3.4: (Impossibility of $\mathsf{Poly}(n)$Query Complexity) If both the following conditions are satisfied:

- $\frac{1}{2}n \geq k_n \geq C_1 n^{\epsilon_1}$
- $\delta_n = \Omega(k_n^{\frac{1+\epsilon_2}{2}})$

where $\epsilon_1, \epsilon_2 \in (0, 1)$, and $C_1 > 0$, then $T_n^*(k_n, \delta_n)$ is $\omega(n^p)$, for all $p \in \mathbb{N}$. In words, there does not exist querying methods with $\mathsf{Poly}(n)$ query complexity that can do the job.

Again, the assumption $\frac{1}{2}n > k_n$ is reasonable due to Remark 3.1. To prove this result, we utilize Chernoff bound to derive a lower bound on $T_n^*(k_n, \delta_n)$, and see that it grows exponentially fast with n if δ_n is great enough. The details can be found in Appendix B of [9].

Remark 3.2: Corollary 3.1 and Theorem 3.4 establish a sharp boundary $\delta_n \approx \sqrt{k_n}$ of partial data extraction under noisy responses to histogram queries. Roughly speaking, if $\delta_n \ll \sqrt{k_n}$, then the sufficient and necessary condition to recover data set is $T_n^* = \Theta(n/\log n)$. On the other hand, if $\delta_n \gg \sqrt{k_n}$, there is no querying method with $\mathsf{Poly}(n)$ query complexity can reconstruct data set successfully.

IV. Achievability via Randomized Querying

In this section, we give the proof of Theorem 3.1. The proof involves upper bounding the probability of failure. Due to the randomized construction of the querying matrix, each entry is generated in an i.i.d. fashion. Therefore, we first cast the probability of failure into the central probability of binomial distribution, and then further upper bound it.

Claim 4.1: Under the randomized query defined in (3), the probability of failure can be upper bounded by

$$\begin{aligned} &P_f(\boldsymbol{x}; k_n, \delta_n) \\ &\leq \sum_{t=k_n}^{n} \binom{n}{t} \mathbb{P}\left(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n\right)^{T_n}, \end{aligned} \tag{6}$$

where $B_t \sim \text{Binomial}(t, 1/2)$.

The proof of the above claim is given in Appendix C in [9].

Continuing the proof of Theorem 3.1, the key is to separate the summation of (6) into two parts:

$$\underbrace{\sum_{t=k_n}^{k^*} \binom{n}{t} \mathbb{P}\left(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n\right)^{T_n} +}_{\text{(i)}}$$

$$\underbrace{\sum_{t=k^*}^{n} \binom{n}{t} \mathbb{P}\left(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n\right)^{T_n}.}_{\text{(ii)}} \tag{7}$$

Before continuing bounding the probability of failure, we give a lemma to upper bound the central probability of binomial distribution:

Lemma 4.1: Let $B_t \overset{iid}{\sim} \text{Binomial}(t, 1/2)$, $\delta_n \in (0, t/16)$ then the following two upper bounds hold:

1) $\mathbb{P}\left(t/2 - 2\delta_n \leq B_t \leq t/2 + 2\delta_n\right) \leq \frac{4\delta_n + 1}{\sqrt{\pi t}}$.

This bound is used when δ_n is small (with respect to t).

2) $\mathbb{P}\left(t/2 - 2\delta_n \le B_t \le t/2 + 2\delta_n\right) \le 1 - \frac{2}{15} e^{-64\delta_n^2/t}$.
This bound is used when δ_n is large (with respect to t).

The proof can be found in Appendix D in [9].

Now, we are ready for upper bounding (7).

For part (i) in (7), applying the second bound in Lemma 4.1, we have

$$\begin{aligned}
&\sum_{t=k_n}^{k^*} \binom{n}{t} \mathbb{P}\left(t/2 - 2\delta_n \le B_t \le t/2 + 2\delta_n\right)^{T_n} \\
\le &\sum_{t=1}^{k^*} \binom{n}{t} \mathbb{P}\left(k_n/2 - 2\delta_n \le B_{k_n} \le k_n/2 + 2\delta_n\right)^{T_n} \\
\le &\sum_{t=1}^{k^*} \binom{n}{t} \left(1 - \frac{2}{15}\exp\left(-\frac{64\delta_n^2}{k_n}\right)\right)^m \\
\le &\left(1 - \frac{2}{15}\exp\left(-\frac{64\delta_n^2}{k_n}\right)\right)^{T_n} (n+1)^{k^*} \qquad (8)
\end{aligned}$$

Due to our assumption that $\delta_n = O(\sqrt{k_n})$, $64\delta_n^2/k_n$ is upper bounded by some constant $\eta \ge 0$ for sufficiently large n, and hence

$$\left(1 - \frac{2}{15}\exp\left(-\frac{64\delta_n^2}{k_n}\right)\right) \le \left(1 - \frac{2}{15}\exp\left(\eta\right)\right) =: \xi,$$

for sufficient large n. Note that ξ is a constant which does not depend on n, and is *strictly* less than 1.

Hence (8) can be further bounded by $\xi^{T_n}(n+1)^{k^*}$. To get vanishing probability of failure, T_n must satisfy

$$T_n = \Omega\left(\frac{k^* \log n}{\log \xi}\right) = \Omega\left(k^* \log n\right), \qquad (9)$$

since ξ does not depend on n.

For part (ii) in (7), we have

$$\begin{aligned}
&\sum_{t=k^*}^{n} \binom{n}{t} \mathbb{P}\left(t/2 - 2\delta_n \le B_t \le t/2 + 2\delta_n\right)^{T_n} \\
\le &\sum_{t=0}^{n} \binom{n}{t} \mathbb{P}\left(k^*/2 - 2\delta_n \le B_{k^*} \le k^*/2 + 2\delta_n\right)^{T_n} \\
\le &\left(\frac{4\delta_n + 1}{\sqrt{\pi k^*}}\right)^{T_n} 2^{n+1}, \qquad (10)
\end{aligned}$$

where (10) is due to Lemma 4.1.

To obtain vanishing failure probability,

$$T_n = \Omega\left(\frac{n+1}{\frac{1}{2}\log(\pi k^*) - \log(4\delta_n + 1)}\right). \qquad (11)$$

Notice that (11) requires $\sqrt{\pi k^*} > 4\delta_n + 1$.

In order to choose a proper k^* according to (9) and (11), we distinguish k_n into two regimes:

1) $k_n = O\left(n^\epsilon\right)$, for some $\epsilon \in (0,1)$:
In this regime, $k_n = O\left(n^\epsilon\right)$ and $\delta_n = O\left(n^{\epsilon/2}\right)$. Hence one can choose k^* such that $k^* = \Theta\left(n^{\epsilon+\epsilon'}\right)$, where $\epsilon + \epsilon' < 1$. In this case,

$$(9) \implies T_n = \Omega\left(n^{\epsilon+\epsilon'} \log n\right)$$

$$(11) \implies T_n = \Omega\left(\frac{n}{(\epsilon+\epsilon')\log n/2 - \log\delta_n}\right)$$

2) $k_n = \omega\left(n^\epsilon\right)$, for all $\epsilon < 1$:
In this regime, $\delta_n = O\left(k_n^{(1-\epsilon')/2}\right)$, and therefore we can choose k^* such that $k^* = \Theta\left(n^{1-\epsilon'}\right)$. In this case,

$$(9) \implies T_n = \Omega\left(n^{1-\epsilon'} \log n\right)$$

$$(11) \implies T_n = \Omega\left(\frac{n}{(1-\epsilon')\log n/2 - \log\delta_n}\right).$$

The proof is complete by noticing that $T_n = \Omega\left(\frac{n}{\log n}\right)$ is sufficient for the cases in the two regimes.

V. Proof of the Combinatorial Lower Bound

In this section, we give the proof of combinatorial lower bound stated in Theorem 3.3.

For notational convenience, let us define the right-hand side of (5) as τ. Then, the theorem is equivalent to the following statement:

For any $T_n \le \tau$, $\exists\, \boldsymbol{x},\ \tilde{\boldsymbol{x}} \in \{0,1\}^n,\ \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| > k_n$, such that $|\mathbf{Q}\boldsymbol{x} - \mathbf{Q}\tilde{\boldsymbol{x}}| \le 2\delta_n$.

The main idea of the proof is as follows. Consider a subset S of all confused pairs $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ differing by at least k_n elements. After each query $\boldsymbol{q}_i$, one can remove some candidates in S according to the response. If for every single query, the number of removed candidates is at most N, then at least $\frac{|S|}{N}$ queries are needed. We will show that $\tau \le \frac{|S|}{N}$. Therefore once $T_n \le \tau$, there exists at least one ambiguous data $\tilde{\boldsymbol{x}}$, and hence the reconstruction is impossible. Moreover, τ is a lower bound of $T_n^*(k_n, \delta_n)$.

For a data set $\boldsymbol{x} \in \{0,1\}^n$, denote an ambiguous data set as $\tilde{\boldsymbol{x}}$. We focus on the collection of all possible pairs of $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ which have the same one norm, and differs from each other exactly k_n's element, that is,

$$\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_1 = k_n, \text{ and } \|\boldsymbol{x}\|_1 = \|\tilde{\boldsymbol{x}}\|_1.$$

Let $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \{0,1\}^n$, and define

$$\begin{aligned}
S_{k_n} &\triangleq \{(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \mid \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_1 = k_n, \|\boldsymbol{x}\|_1 = \|\tilde{\boldsymbol{x}}\|_1\} \\
&= \left\{(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \mid \pi(1|\boldsymbol{x} - \tilde{\boldsymbol{x}}) = \pi(-1|\boldsymbol{x} - \tilde{\boldsymbol{x}}) = \frac{k_n}{2}\right\},
\end{aligned}$$

where we use $\pi(\cdot \mid \boldsymbol{w})$ to denote the *unnormalized histogram* of vector $\boldsymbol{w}$, say, $\pi(x \mid \boldsymbol{w}) \triangleq$ (number of x in $\boldsymbol{w}$). Define the collection of all *confusion datasets* after the i-th query :

$$V_i \triangleq \{\, (\boldsymbol{x}, \tilde{\boldsymbol{x}}) \in S_{k_n} \mid |\boldsymbol{q}_i \cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}})| \le 2\delta_n \,\}.$$

As long as

$$T_n < \frac{|S_{k_n}|}{\max_{i\in\{1,\dots,T_n\}} |V_i^c|}, \qquad (12)$$

(with a slight abuse of notation, let $V_i^c = V_i^c \cap S_{k_n}$), we have

$$|S_{k_n}| > T_n \max_{i\in\{1,\dots,T_n\}} |V_i^c| \ge \sum_{i=1}^{T_n} |V_i^c| \ge \left|\bigcup_{i=1}^{T_n} V_i^c\right|$$

due to union bound. Notice that

$$|\bigcup_{i=1}^{T_n} V_i^c| < |S_{k_n}| \iff \bigcup_{i=1}^{T_n} V_i^c \neq S_{k_n} \iff \bigcap_{i=1}^{T_n} V_i \neq \emptyset,$$

which implies that there exists at least one pair of confusion data sets $(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \in S_{k_n}$ after T_n independent queries. To complete the proof, all we need is to the following claim:

Claim 5.1:

$$\tau \leq \frac{|S_{k_n}|}{\max_{i\in\{1,...,T_n\}} |V_i^c|}.$$

Proof. First, we introduce

$$\mathcal{T}_1 \triangleq \{j \,|\tilde{x}_j = 0, x_j = 1\},\ \mathcal{T}_2 \triangleq \{j \,|\tilde{x}_j = 1, x_j = 0\}.$$

Note that suppose $(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \in S_{k_n}$, then $|\mathcal{T}_1| = |\mathcal{T}_2| = k_n/2$ due to the fact $\|\boldsymbol{x}\|_1 = \|\tilde{\boldsymbol{x}}\|_1$, and $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_1 = k_n$. Then obviously we have

$$|S_{k_n}| = \binom{n}{k_n/2}\binom{n-k_n/2}{k_n/2} 2^{n-k_n}. \tag{13}$$

Let the queried subset corresponding to $\boldsymbol{q}$ be $\mathcal{S}$. The *confusion events* $\{|\boldsymbol{q}\boldsymbol{x} - \boldsymbol{q}\tilde{\boldsymbol{x}}| \leq 2\delta_n\}$ happen if and only if

$$\big||\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2|\big| \leq 2\delta_n. \tag{14}$$

Therefore, to upper bound $|V_i^c|$, we have

$$\begin{aligned}&\max_i |V_i^c| \leq \max_{\boldsymbol{q}} |\,\{\,(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \in S_{k_n} \ |\, |\boldsymbol{q} \cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}})| > 2\delta_n\,\}\,| \\ &= \max_{\mathcal{S} \subset [n]} \big|\, \{(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \in S_{k_n} \,|||\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2|| > 2\delta_n\} \,\big|\end{aligned} \tag{15}$$

By symmetry, it is intuitive that the maximum is attained when $|\mathcal{S}| = \frac{n}{2}$ (we also give a rigorous proof in Appendix, see Lemma D.1), and thus (15) is equal to

$$\begin{aligned}&\big|\,\{(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \in S_{k_n} \,|||\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2|| > 2\delta_n\} \\ =&\left|\bigcup_{\delta > 2\delta_n} \{(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \in S_{k_n} \mid ||\mathcal{S} \cap \mathcal{T}_1| - |\mathcal{S} \cap \mathcal{T}_2|| = \delta\}\right| \\ =&2^{n-k_n+1} \sum_{\alpha=0}^{k_n/2} \sum_{\delta=2\delta_n}^{\alpha} \binom{n/2}{\alpha}\binom{n/2}{k_n/2-\alpha} \times \\ &\binom{n/2-\alpha}{\alpha-\delta}\binom{n/2-k_n/2+\alpha}{k_n/2+\delta-\alpha}\end{aligned} \tag{16}$$

Combining (13) and (16), we obtain

$$\begin{aligned}&\frac{|S_{k_n}|}{\max_{i\in\{1,...,T_n\}} |V_i^c|} \\ \geq&\frac{\binom{n}{k_n/2}\binom{n-k_n/2}{k_n/2}}{2\sum_{\alpha=0}^{k_n/2}\sum_{\delta=2\delta_n}^{\alpha}\binom{n/2}{\alpha}\binom{n/2-k_n/2}{k_n/2-\alpha}\binom{n/2-\alpha}{\alpha-\delta}\binom{n/2-k_n/2+\alpha}{k_n/2+\delta-\alpha}} \\ =&\frac{\binom{n}{\frac{n}{2}}}{2\sum_{\alpha=2\delta_n}^{k_n/2}\binom{k_n/2}{\alpha}\sum_{\delta=2\delta_n}^{\alpha}\binom{k_n/2}{\alpha-\delta}\binom{n-k_n}{n/2-2\alpha+\delta}} = \tau,\end{aligned} \tag{17}$$

where (17) is due to direct calculation of binomial coefficient. This proves our claim. ■

VI. Extension

We close this paper by briefly explaining how to extend our results to the general case $|\mathcal{A}| = d$. Following the formulation in [2], the data set can be modeled by a matrix $\mathbf{X} \in \{0,1\}^{n\times d}$, and the response $\mathbf{Y} \in \{0, 1, ..., n\}^{T_n \times d}$. To prove the achievability part, we notice that the probability of error $P_f(\mathbf{X}; k_n, \delta_n)$ (w.r.t $\mathbf{Q}$) is upper bounded by $P_f(\boldsymbol{x}; k_n, \delta_n)$; here we abuse the notation, denoting $\boldsymbol{x} \in \{0,1\}^n$ for some column of $\mathbf{X}$. Hence, Theorem 3.1 also holds for d being constant with respect to n.

On the other hand, suppose $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are two data sets with Hamming distance greater than k_n. Then, there exists some column of $\mathbf{X}, \tilde{\mathbf{X}}$, say, $\boldsymbol{x}, \tilde{\boldsymbol{x}}$, such that $d_{\text{data}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \geq k_n/d$. Therefore, the converse results in Therem 3.2, 3.3, and 3.4 hold for $k'_n = \left(\frac{k_n}{d}\right)$. In particular, as long as d is a constant with respect to n, the asymptotic behavior remains the same.

Acknowledgment

The work was supported by the Ministry of Science and Technology, Taiwan, under Grant 105-2221-E-002-018.

References

[1] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Theoretical Computer Science*, pp. 211–407, 2013.

[2] I.-H. Wang, S.-L. Huang, K.-Y. Lee, and K.-C. Chen, “Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms,” *Proceedings of IEEE International Symposium on Information Theory*, July 2016.

[3] I.-H. Wang, S.-L. Huang, and K.-Y. Lee, “Extracting sparse data via histogram queries,” *Proceedings of Annual Allerton Conference on Communications, Control, and Computing*, September 2016.

[4] A. E. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborova, and M. I. Jordan, “Decoding from pooled data: Sharp information-theoretic bounds,” *arXiv:1611.09981*, 2016.

[5] I. Dinur and K. Nissim, “Revealing information while preserving privacy,” *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 202–210, 2003.

[6] C. Dwork and S. Yekhanin, “New efficirnt attacks on statistical disclosure control mechanism,” *In Proceedings of the Advances in Cryptology-CRYPTO*, pp. 469–480, 2008.

[7] C. Dwork, F. McSherry, and K. Talwar, “The price of privacy and the limits of lp decoding,” *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 85–94, June 2007.

[8] M. Bun, J. Ullman, and S. Vadhan, “Fingerprinting codes and the price of approximate differential privacy,” *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, 2014.

[9] W.-N. Chen and I.-H. Wang, “Partial data extraction via noisy histogram queries: Information theoretic bounds,” *Manuscript*, January 2017. http://homepage.ntu.edu.tw/%7Eihwang/Research/Eprints/isit17nq.pdf.

[10] J. Matoušek and J. Vondrák, “The probabilistic method,” *Lecture Notes*, March 2008.

Appendix A
proof of Theorem 3.2

Theorem 3.2: (Packing Lower Bound) Let $k_n \le \left(\frac{1-\epsilon}{2}\right) n$ for some $\epsilon > 0$. Then, the following lower bound holds:

$$T_n^*(k_n, \delta_n) = \Omega\left(\frac{n\left(1 - H_b\left(\frac{1-\epsilon}{2}\right)\right)}{\log(n+1) - \log(4\delta_n + 1)}\right) \tag{4}$$

Specifically, when $\delta_n = O(n^{\frac{1-\epsilon'}{2}})$, and ϵ, ϵ' does not depend on n, then (4) can be further simplified to

$$T_n^*(k_n, \delta_n) = \Omega\left(n/\log n\right).$$

We prove Theorem 3.2 by packing argument.

Proof. (proof of Theorem 3.2)
To reconstruct $\boldsymbol{x}$ successfully, the cardinality of possible input must less than the cardinality of possible output. The number of possible input (output) turns out to be a packing problem. First, we notice the number of $\tilde{\boldsymbol{x}}$ with distance to $\boldsymbol{x}$ less than k_n is

$$\left|\{\tilde{\boldsymbol{x}} \in \{0,1\}^n \mid \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_1 \le k_n\}\right| = \sum_{i=1}^{k_n} \binom{n}{i},$$

thus the total number of all possible pairs $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ with distance greater than k_n to each other is

$$\left|\{\boldsymbol{x_i} \in \{0,1\}^n \mid |\boldsymbol{x_i} - \boldsymbol{x_j}| > k_n, \forall \boldsymbol{x_i} \ne \boldsymbol{x_j}\}\right| = \frac{2^n}{\sum_{i=1}^{k_n} \binom{n}{i}}.$$

On the other hand, the total number of possible outcomes is

$$\frac{\left|\{0,1,...,n\}^{T_n}\right|}{(4\delta_n + 1)^{T_n}} = \left(\frac{n+1}{4\delta_n + 1}\right)^{T_n}$$

To guarantee successful reconstruction, the number of possible outputs must greater than the number of possible input, hence we have

$$\frac{2^n}{\sum_{i=1}^{k_n} \binom{n}{i}} \le \left(\frac{n+1}{4\delta_n + 1}\right)^{T_n} \tag{18}$$

Now, we give a bound on summation of binomial coefficient: Suppose $k = pn$, where $p \in (0,1)$ does not depend on n. Then by Stirling approximation, we have

$$\log\binom{n}{pn} = nH_b\left(p\right) + O(\log n),$$

where $H_b(p)$ is the binary entropy function.

$$\begin{aligned} T_n &\ge \frac{n - \log\left(\sum_{i=1}^{k_n} \binom{n}{i}\right)}{\log(n+1) - \log(4\delta_n + 1)} && (19) \\ &\ge \frac{n - \log\left(k_n \binom{n}{n(1-\epsilon)/2}\right)}{\log(n+1) - \log(4\delta_n + 1)} && (20) \\ &= \frac{n\left(1 - H_b\left(\frac{1-\epsilon}{2}\right) + O(\frac{\log n}{n})\right)}{\log(n+1) - \log(4\delta_n + 1)}, && (21) \end{aligned}$$

where (20) is due to $\sum_{i=1}^{k_n} \binom{n}{i} \le k_n \binom{n}{k_n}$ and $k_n \le \left(\frac{1-\epsilon}{2}\right) n$.
Furthermore, if $\delta_n = O(n^{\frac{1-\epsilon'}{2}})$, and ϵ, ϵ' does not depend on n, then we have

$$T_n = \Omega\left(\frac{n}{\log n}\right).$$

■

Appendix B
proof of Theorem 3.4

Theorem 3.4: (Impossibility of Poly(n)Query Complexity)
If both the following conditions are satisfied:

- $\frac{1}{2}n \geq k_n \geq C_1 n^{\epsilon_1}$
- $\delta_n = \Omega(k_n^{\frac{1+\epsilon_2}{2}})$

where $\epsilon_1, \epsilon_2 \in (0,1)$, and $C_1 > 0$, then $T_n^*(k_n, \delta_n)$ is $\omega(n^p)$, for all $p \in \mathbb{N}$. In words, there does not exist querying methods with Poly(n) query complexity that can do the job.

Before we further bounding (5), we give two technical lemma:

Lemma B.1: For $n \geq 2$, the following binomial bound holds:

$$\frac{4^n}{\sqrt{\frac{\pi}{2}(2n+1)}} \leq \binom{2n}{n} \leq \frac{4^n}{\sqrt{\pi n}}$$

Lemma B.2: For $\delta \leq n/2$, the following bound holds:

$$\sum_{k=n/2-\delta}^{n/2+\delta} \binom{n}{k} \geq 2^n \left(1 - 2\exp\left(-\frac{\delta^2}{n}\right)\right)$$

Now, we are ready to prove Theorem 3.4.

Proof. (proof of Theorem 3.4)

We show that as long as $k_n = \Omega\left(n^{\epsilon}\right)$ and $\delta_n = \Omega\left(k_n^{\frac{1+\epsilon'}{2}}\right)$, the bound given in Theorem 3.3 is $\Omega(n)$.

From Theorem 3.3, we know that as long as

$$T_n \leq \tau = \frac{\binom{n}{\frac{n}{2}}}{2\sum_{\alpha=2\delta_n}^{k_n/2}\binom{k_n/2}{\alpha}\sum_{\delta=2\delta_n}^{k_n/2}\binom{k_n/2}{\alpha-\delta}\binom{n-k_n}{n/2-2\alpha+\delta}}, \tag{22}$$

the data set cannot be reconstructed successfully.

Applying Lemma B.1, we see that

1) $\binom{n}{n/2} \geq \frac{2^n}{\sqrt{\frac{\pi}{2}(n+1)}}$
2) $\binom{n-k_n}{n/2-2\alpha+\delta} \leq \binom{n-k_n}{(n-k_n)/2} \leq \frac{2^{n-k_n}}{\sqrt{\frac{\pi}{2}(n-k_n)}}$.

Thus (22) is lower bounded by

$$\begin{aligned}
&\geq \frac{2^{n-1}}{\sqrt{\frac{\pi}{2}(n+1)}} \times \\
&\left\{\sum_{\alpha=2\delta_n}^{k_n/2}\binom{k_n/2}{\alpha}\sum_{\delta=2\delta_n}^{\alpha}\binom{k_n/2}{\alpha-\delta}\left(\frac{2^{n-k_n}}{\sqrt{\frac{\pi}{2}(n-k_n)}}\right)\right\}^{-1} \\
&= \frac{\sqrt{n-k_n}}{2\sqrt{n+1}}\left\{2^{-k_n}\sum_{\alpha=2\delta_n}^{k_n/2}\binom{k_n/2}{\alpha}\sum_{\delta=2\delta_n}^{\alpha}\binom{k_n/2}{\alpha-\delta}\right\}^{-1} \\
&= \frac{\sqrt{n-k_n}}{2\sqrt{n+1}}\left\{2^{-k_n}\sum_{\alpha=2\delta_n}^{k_n/2}\binom{k_n/2}{\alpha}\sum_{i=0}^{\alpha-2\delta_n}\binom{k_n/2}{i}\right\}^{-1} \qquad (23)
\end{aligned}$$

Notice that

$$\begin{aligned}
&\sum_{\alpha=2\delta_n}^{k_n/2}\binom{k_n/2}{\alpha}\sum_{i=0}^{\alpha-2\delta_n}\binom{k_n/2}{i} \\
&\leq 2^{k_n} - \left(\sum_{j=k_n/4-2\delta_n}^{k_n/4+2\delta_n}\binom{k_n/2}{j}\right)^2 \qquad (24) \\
&\leq 2^{k_n} - \left[2^{k_n/2}\left(1 - 2\exp\left(-\frac{2\delta_n^2}{k_n}\right)\right)\right]^2 \qquad (25)
\end{aligned}$$

$$= 4\exp\left(-\frac{2\delta_n^2}{k_n}\right) - 4\exp\left(-\frac{2\delta_n^2}{k_n}\right)^2, \tag{26}$$

where (24) is due to the observation of summation region, and (25) is due to Lemma (B.2).
Applying (26), we have

$$\begin{aligned}(23) \geq & \frac{\sqrt{n-k_n}}{8\sqrt{n+1}}\left\{\exp(-\frac{2\delta_n^2}{k_n}) - \exp\left(-\frac{2\delta_n^2}{k_n}\right)^2\right\}^{-1} \\ = & \underbrace{\frac{\sqrt{n-k_n}}{8\sqrt{n+1}}}_{\text{(a)}}\underbrace{\exp\left(\frac{2\delta_n^2}{k_n}\right)}_{\text{(b)}}\underbrace{\left\{1-\exp\left(-\frac{2\delta_n^2}{k_n}\right)\right\}^{-1}}_{\text{(c)}}\end{aligned}$$

As long as $n \to \infty$,
1) $(a) \geq \frac{1}{8\sqrt{2}}$, due to the assumption $\frac{1}{2}n > k_n$
2) $(b) = \omega(n^p)$ for all integer p, due to the fact

$$\begin{aligned}\exp\left(\frac{2\delta_n^2}{k_n}\right) &\geq \exp\left(k_n^{\epsilon_2/2}\right) \geq \exp\left(C_1 n^{\epsilon_1\epsilon_2/2}\right) \\ &\geq \exp\left(p\log n\right) = n^p\end{aligned}$$

3) $(c) \geq 1$
Combine (a), (b), (c) together, we conclude that as long as T_n polynomial in n, the successful recovery is impossible. ■

Appendix C
Proof of Claim 4.1

First, we use the notation $B_{n_1}^{(1)}, B_{n_2}^{(2)}$ to denote the independent random variables with distribution $\mathrm{Binomial}(n_1, 1/2)$ and $\mathrm{Binomial}(n_2, 1/2)$ respectively. By the definition of probability of failure, $P_f(\boldsymbol{x}; k_n, \delta_n)$ is

$$P_Q\left(\exists\tilde{\boldsymbol{x}}, \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_1 > k_n, \|\mathbf{Q}\tilde{\boldsymbol{x}} - \mathbf{Q}x\|_\infty \leq 2\delta_n\right) \tag{27}$$

$$= P_Q\left(\bigcup_{\tilde{\boldsymbol{x}}\in B_{k_n}^c(\boldsymbol{x})} \|\mathbf{Q}\tilde{\boldsymbol{x}} - \mathbf{Q}x\|_\infty \leq 2\delta_n\right) \tag{28}$$

$$\leq \sum_{\tilde{\boldsymbol{x}}\in B_{k_n}^c(\boldsymbol{x})} \mathbb{P}\left(|\boldsymbol{q}\boldsymbol{x} - \boldsymbol{q}\tilde{\boldsymbol{x}}| \leq 2\delta_n\right)^{T_n} \tag{29}$$

$$= \sum_{t=k_n}^{n}\sum_{\tilde{\boldsymbol{x}}\in\partial B_t(\boldsymbol{x})} \mathbb{P}\left(|\boldsymbol{q}\boldsymbol{x} - \boldsymbol{q}\tilde{\boldsymbol{x}}| \leq 2\delta_n\right)^{T_n} \tag{30}$$

$$\leq \sum_{t=k_n}^{n}\binom{n}{t}\max_{t^+ + t^- = t}\mathbb{P}\left(\left|B_{t^+}^{(1)} - B_{t^-}^{(2)}\right| \leq 2\delta_n\right)^{T_n} \tag{31}$$

Here we use $B_R(\boldsymbol{x})$ to denote the ball centered at $\boldsymbol{x}$ with radius R, and use $\partial B_R(\boldsymbol{x})$ to denote the boundary of $B_R(\boldsymbol{x})$. Notice that (29) is due to union bound, (31) is due to the fact that each $\boldsymbol{q}_i$ is generated according to $\mathrm{Ber}(1/2)$. To handle (31), we give the following lemma:

Lemma C.1: For $t_1 + t_2 = T$, T is even, the following fact holds:

$$\begin{aligned}&\mathbb{P}\left(\left|B_{t_1}^{(1)} - B_{t_2}^{(2)}\right| \leq \delta\right) \\ \leq &\mathbb{P}\left(\left|B_{T/2}^{(1)} - B_{T/2}^{(2)}\right| \leq \delta\right),\end{aligned}$$

where $B_{t_1}^{(1)}, B_{t_2}^{(2)}$ are independent random variables with distribution $\mathrm{Binomial}(n_1, 1/2)$ and $\mathrm{Binomial}(n_2, 1/2)$ respectively.

From lemma C.1, we see that the maximum of (31) occurs when $t^+ = t^- = t/2$. For simplicity we assume t even, and (31) becomes

$$\sum_{t=k_n}^{n}\binom{n}{t}\mathbb{P}\left(\left|B_{t/2}^{(1)} - B_{t/22}^{(2)}\right| \leq 2\delta_n\right)^{T_n} \tag{32}$$

$$= \sum_{t=k_n}^{n}\binom{n}{t}\mathbb{P}\left(\left|B_{t/2}^{(1)} - B_{t/2}^{(2)} - \frac{t}{2}\right| \leq 2\delta_n\right)^{T_n} \tag{33}$$

$$=\sum_{t=k_n}^{n}\binom{n}{t}\mathbb{P}\left(\left|B_t-\frac{t}{2}\right|\le 2\delta_n\right)^{T_n} \tag{34}$$

$$=\sum_{t=k_n}^{n}\binom{n}{t}\mathbb{P}\left(t/2-2\delta_n\le B_t\le t/2+2\delta_n\right)^{T_n}, \tag{35}$$

here B_t in (33) denotes the random variable with distribution $\mathrm{Binomial}(t,1/2)$. (33) is due to the basic combinatorial fact, and (34) is due to the fact that our construction of $\mathbf{Q}$ is independent.

Appendix D
Technical Lemmas

A. Lemma D.1

Lemma D.1: Let $S_{k_n}, V_i, \mathcal{T}_1$ and $\mathcal{T}_2$ be defined as before. Then

$$\max_{\mathcal{S}\subset[n]}\left|\left\{(\boldsymbol{x},\tilde{\boldsymbol{x}})\in S_{k_n}\;\middle|\;||\mathcal{S}\cap\mathcal{T}_1|-|\mathcal{S}\cap\mathcal{T}_2||>2\delta_n\right\}\right| \tag{36}$$

achieves its maximum when $|\mathcal{S}|=\frac{n}{2}$.

Proof. First, let $|S|=s$, $s\in[n]$. Then (36) becomes

$$\begin{aligned}&2^{n-k_n+1}\sum_{\alpha=0}^{k_n/2}\sum_{\delta=2\delta_n}^{\alpha}\binom{s}{\alpha}\binom{n-s}{k_n/2-\alpha}\binom{s-\alpha}{\alpha-\delta}\binom{n-s-k_n/2+\alpha}{k_n/2+\delta-\alpha}\\ =&2^{n-k_n+1}\binom{n}{k_n/2,k_n/2,n-k_n}\frac{\sum_{\alpha=0}^{k_n/2}\sum_{\delta=2\delta_n}^{\alpha}\binom{k_n/2}{\alpha}\binom{k_n/2}{\alpha-\delta}\binom{n-k_n}{s-2\alpha+\delta}}{\binom{n}{s}}\end{aligned} \tag{37}$$

Therefore, maximize (36) is equivalent to maximize (37) over all possible s. After change of variables, (37) becomes

$$\begin{aligned}&\frac{\sum_{|i-j|>2\delta_n}\binom{k_n/2}{i}\binom{k_n/2}{j}\binom{n-k_n}{s-(i+j)}}{\binom{n}{s}}\\ =&\frac{\sum_k\sum_{\substack{i+j=k\\|i-j|>2\delta_n}}\binom{k_n/2}{i}\binom{k_n/2}{j}\binom{n-k_n}{s-k}}{\binom{n}{s}}\\ =&\sum_k\underbrace{\left(\frac{\sum_{\substack{i+j=k\\|i-j|>2\delta_n}}\binom{k_n/2}{i}\binom{k_n/2}{j}}{\binom{k_n}{k}}\right)}_{a_k}\underbrace{\left(\frac{\binom{k_n}{k}\binom{n-k_n}{s-k}}{\binom{n}{s}}\right)}_{b_k(s)}\\ =&\sum_k a_k\cdot b_k(s).\end{aligned}$$

One can observe the following facts:

- a_k is symmetric to $k=k_n/2$, that is, $a_k=a_{k_n-k}$, since one can change the variables $(i',j')=(k_n/2-i,k_n/2-j)$.
- a_k is maximized as $k=k_n/2$. This can be proved by writing a_k in another form:

  $$a_k=\frac{\sum_{\substack{i>(k+2\delta_n)/2\text{ or }\\ i<(k-\delta_n)/2}}\binom{k}{i}\binom{k_n-k}{k_n/2-i}}{\binom{k_n}{k_n/2}},$$

  and it achieves maximum at $k=k_n/2$. Also, a_k is increasing in $[0,k_n/2]$ (and hence decreasing in $[k_n/2,k_n]$).
- For all $s\in[n]$, $\sum_k b_k(s)=1$.
- For $s\in[0,n/2)$, $b_k(s)$ is maximized at $k^*\in[0,k_n/2)$; for $s\in(n/2,n]$, $b_k(s)$ is maximized at $k^*\in(k_n/2,k_n]$. Also, $b_k(s)$ is increasing for $k\le k^*$, and decreasing for $k\ge k^*$.
- $b_k(s)$ is symmetric to $(n/2,k_n/2)$, that is,

  $$b_k(s)=b_{k_n-k}(n-s).$$

- $\sum_k a_k\cdot b_k(s)$ is symmetric to $n/2$, that is, $\sum_k a_k\cdot b_k(s)=\sum_k a_k\cdot b_k(n-s)$.

Now, we are ready to show $\sum_k a_k \cdot b_k(s)$ attains its maximum at $\boldsymbol{x} = n/2$. For any $s \in [n]$, we have

$$\sum_k a_k \cdot b_k(s) = \sum_k a_k \cdot \left(\frac{b_k(s) + b_k(n-s)}{2}\right).$$

Consider the equation

$$b_k(n/2) = \left(\frac{b_k(s) + b_k(n-s)}{2}\right).$$

There's exactly a zero at $k = \xi \in [0, n/2]$ for all s, and due to symmetry, there's another zero at $k = k_n - \xi$. Besides, $b_k(n/2) \geq \left(\frac{b_k(s)+b_k(n-s)}{2}\right)$ at $k = k_n/2$. Therefore we conclude that

$$b_k(n/2) \geq \left(\frac{b_k(s) + b_k(n-s)}{2}\right)$$

for $s \in [\xi, k_n - \xi]$, and

$$b_k(n/2) < \left(\frac{b_k(s) + b_k(n-s)}{2}\right)$$

for $s \in [\xi, k_n - \xi]^c$ (With a slight abuse of notation, we denote $[0, k_n]\backslash[\xi, k_n - \xi]$ as $[\xi, k_n - \xi]^c$).
Notice that since

$$\sum_k b_k(s) = \sum_k \left(\frac{b_k(s) + b_k(n-s)}{2}\right) = 1,$$

we have

$$\begin{aligned}
&\sum_{k \in [\xi, k_n - \xi]} \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2}\right)\right) \\
=& - \sum_{k \in [\xi, k_n - \xi]^c} \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2}\right)\right).
\end{aligned} \tag{38}$$

Now, consider

$$\begin{aligned}
&\sum_k a_k \cdot b_k(n/2) - \sum_k a_k \cdot b_k(s) \\
=&\sum_k a_k \cdot \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2}\right)\right) \\
=& \sum_{k \in [\xi, k_n - \xi]} a_k \cdot \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2}\right)\right) \\
&+ \sum_{k \in [\xi, k_n - \xi]^c} a_k \cdot \left(b_k(n/2) - \left(\frac{b_k(s) + b_k(n-s)}{2}\right)\right) \\
\geq& 0.
\end{aligned} \tag{39}$$

(39) is due to the fact that $a_{k_1} \geq a_{k_2}$, for all $k_1 \in [\xi, k_n - \xi]$, $k_2 \in [\xi, k_n - \xi]^c$ and (38). Since it holds for all $s \in [n]$, the proof is complete. ■

B. Proof of Lemma C.1

Lemma C.1: For $t_1 + t_2 = T$, T is even, the following fact holds:

$$\begin{aligned}
&\mathbb{P}\left(\left|B_{t_1}^{(1)} - B_{t_2}^{(2)}\right| \leq \delta\right) \\
\leq&\mathbb{P}\left(\left|B_{T/2}^{(1)} - B_{T/2}^{(2)}\right| \leq \delta\right),
\end{aligned}$$

where $B_{t_1}^{(1)}, B_{t_2}^{(2)}$ are independent random variables with distribution $\text{Binomial}(n_1, 1/2)$ and $\text{Binomial}(n_2, 1/2)$ respectively.

Proof. First, note that for $B_t \sim \text{Binomial}\left(t, \frac{1}{2}\right)$, $t - B_t$ has the same distribution with B_t. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\left|B_{t_1}^{(1)} - B_{t_2}^{(2)}\right| \leq \delta\right) \\
=&\mathbb{P}\left(\left|B_{t_1+t_2} - t_2\right| \leq \delta\right)
\end{aligned}$$

$$
\begin{aligned}
&=\mathbb{P}\left(t_2-\delta \le B_T \le t_2+\delta\right)\\
&\le\mathbb{P}\left(T/2-\delta \le B_T \le T/2+\delta\right)\\
&=\mathbb{P}\left(\left|B^{(1)}_{T/2}-B^{(2)}_{T/2}\right| \le \delta\right).
\end{aligned}
$$

■

C. *Proof of Lemma 4.1*

Lemma 4.1: Let $B_t \overset{iid}{\sim}$ Binomial$(t, 1/2)$, $\delta_n \in (0, t/16)$ then the following two upper bounds hold:

1) $\mathbb{P}\left(t/2 - 2\delta_n \le B_t \le t/2 + 2\delta_n\right) \le \frac{4\delta_n+1}{\sqrt{\pi t}}$.
   This bound is used when δ_n is small (with respect to t).
2) $\mathbb{P}\left(t/2 - 2\delta_n \le B_t \le t/2 + 2\delta_n\right) \le 1 - \frac{2}{15}e^{-64\delta_n^2/t}$.
   This bound is used when δ_n is large (with respect to t).

The proof can be found in Appendix D in [9].

Proof.

1) Since $\mathbb{P}(B_t = t/2) > \mathbb{P}(B_t = k)$, for all $k \in [0, t]$, we have

$$
\begin{aligned}
\mathbb{P}\left(t/2 - 2\delta_n \le B_t \le t/2 + 2\delta_n\right) \le&(4\delta_n + 1)\mathbb{P}(B_t = t/2)\\
\le&\frac{4\delta_n + 1}{\sqrt{\pi t}}.
\end{aligned}
$$

   The last inequality is due to Lemma B.1.

2) For convenience, let $\delta_n = 2\delta_n'$. This is equivalent to show

$$
\mathbb{P}(B_t > t/2 + \delta_n) \ge \frac{1}{15}\exp\left(-16\delta_n^2/t\right).
$$

   The proof is first given in [10], which involves some elementary estimates. For the sake of completeness, we state it again. Write $t = 2m$. We have

$$
\begin{aligned}
&\mathbb{P}(B_t \ge m + \delta_n) = 2^{-2m}\sum_{j=\delta_n}^{m}\binom{2m}{m+j}\\
&\ge 2^{-2m}\sum_{j=\delta_n}^{2\delta_n-1}\binom{2m}{m+j}\\
&=2^{-2m}\sum_{j=\delta_n}^{s\delta_n-1}\binom{2m}{m}\frac{m}{m+j}\cdot\frac{m-1}{m+j-1}\cdots\frac{m-j+1}{m+1}\\
&\ge\frac{1}{\sqrt{m}}\sum_{j=\delta_n}^{2\delta_n-1}\prod_{i=1}^{j}\left(1-\frac{j}{m+i}\right)\\
&\ge\frac{1}{\sqrt{m}}\left(1-\frac{2\delta_n}{m}\right)^{2\delta}\\
&\ge\frac{1}{\sqrt{m}}\cdot\exp(-8\delta_n^2/m).
\end{aligned}
$$

   For $\delta_n \ge \frac{\sqrt{m}}{4}$, the last expression is at least $\frac{1}{8}\exp\left(\frac{-16\delta_n}{n}\right)$. For $0 \le \delta_n < \frac{\sqrt{m}}{4}$, we have

$$
\mathbb{P}(B_t > m + \delta_n) > \mathbb{P}(B_t > m + \frac{\sqrt{m}}{4}) \ge \frac{1}{8}\exp(-1/2) > \frac{1}{15}.
$$

   Thus the claimed bound holds for all $\delta_n \le m/4$.

■

D. *Proof of Lemma B.1*

Lemma B.1: For $n \geq 2$, the following binomial bound holds:

$$\frac{4^n}{\sqrt{\frac{\pi}{2}(2n+1)}} \leq \binom{2n}{n} \leq \frac{4^n}{\sqrt{\pi n}}$$

Proof. First we consider the two expressions:

$$2n\left(\binom{2n}{n}\frac{1}{4^n}\right)^2 = \underbrace{\frac{1}{2}\frac{3}{2}\frac{3}{4}\frac{5}{4}\cdots\frac{2n-1}{2n-2}}_{(1)}\underbrace{\frac{2n-1}{2n}}_{(2)} \tag{40}$$

$$=\frac{1}{2}\prod_{j=2}^{n}\left(1+\frac{1}{4j(j-1)}\right), \tag{41}$$

$$(2n+1)\left(\binom{2n}{n}\frac{1}{4^n}\right)^2 \tag{42}$$

$$=\underbrace{\frac{1}{2}\frac{3}{2}\frac{3}{4}\frac{5}{4}\cdots\frac{2n-1}{2n-2}\frac{2n-1}{2n}}_{(1)}\underbrace{\frac{2n+1}{2n}}_{(3)} = \prod_{j=1}^{n}\left(1-\frac{1}{4j^2}\right). \tag{43}$$

By Wallis's formula, (1) converges to $\frac{2}{\pi}$, and (2), (3) converge to 1. Therefore, both (41), (43) converge to $\frac{2}{\pi}$. Notice that according to the left hand side of two expressions, (41) is increasing and (43) is decreasing, with the same limit. Therefore we conclude that

$$2n\left(\binom{2n}{n}\frac{1}{4^n}\right)^2 \leq \frac{2}{\pi}, \tag{44}$$

and

$$(2n+1)\left(\binom{2n}{n}\frac{1}{4^n}\right)^2 \geq \frac{2}{\pi}. \tag{45}$$

Since this holds for $n \geq 2$, the proof is complete. ■

E. *Proof of Lemma B.2*

Lemma B.2: For $\delta \leq n/2$, the following bound holds:

$$\sum_{k=n/2-\delta}^{n/2+\delta}\binom{n}{k} \geq 2^n\left(1-2\exp\left(-\frac{\delta^2}{n}\right)\right)$$

Proof. This is a direct application of Chernoff Bound. Let $X_i \overset{i.i.d}{\sim} \mathrm{Ber}(\frac{1}{2})$, $i \in [N]$. Applying Chernoff inequality on $X = \sum_1^N X_i$, we have

$$\mathbb{P}(X \geq \mathbb{E}X + \delta) \leq \exp\left(\frac{-\delta^2}{n}\right).$$

Therefore,

$$\begin{aligned}\sum_{k=n/2-\delta}^{n/2+\delta}\binom{n}{k} \geq& 2^n\left(1-2\exp\left(-\frac{2\delta^2}{n}\right)\right)\\ \geq& 2^n\left(1-2\exp\left(-\frac{\delta^2}{n}\right)\right)\end{aligned}$$

■