8

Lower Bounds and Separation Results

In this section, we investigate various lower bounds and tradeoffs:

1. How *inaccurate* must responses be in order not to completely destroy any reasonable notion of privacy?
2. How does the answer to the previous question depend on the number of queries?
3. Can we separate $(\varepsilon, 0)$-differential privacy from (ε, δ)-differential privacy in terms of the accuracy each permits?
4. Is there an intrinsic difference between what can be achieved for linear queries and for arbitrary low-sensitivity queries while maintaining $(\varepsilon, 0)$-differential privacy?

A different flavor of separation result distinguishes the computational complexity of generating a *data structure* handling all the queries in a given class from that of generating a *synthetic database* that achieves the same goal. We postpone a discussion of this result to Section 9.

8.1 Reconstruction attacks

We argued in Section 1 that any non-trivial mechanism must be randomized. It follows that, at least for some database, query, and choice of random bits, the response produced by the mechanism is not perfectly accurate. The question of how *in*accurate answers must be in order to protect privacy makes sense in all computational models: interactive, non-interactive, and the models discussed in Section 12.

For the lower bounds on distortion, we assume for simplicity that the database consists of a single — but very sensitive — bit per person, so we can think of the database as an n-bit Boolean vector $d = (d_1, \ldots, d_n)$. This is an abstraction of a setting in which the database rows are quite complex, for example, they may be medical records, but the attacker is interested in one specific field, such as the presence or absence of the sickle cell trait. The abstracted attack consists of issuing a string of queries, each described by a subset S of the database rows. The query is asking how many 1's are in the selected rows. Representing the query as the n-bit characteristic vector $\mathbf{S}$ of the set S, with 1s in all the positions corresponding to rows in S and 0s everywhere else, the true answer to the query is the inner product $A(S) = \sum_{i=1}^{n} d_i \mathbf{S}_i$.

Fix an arbitrary privacy mechanism. We will let $r(S)$ denote the response to the query S. This may be obtained explicitly, say, if the mechanism is interactive and the query S is issued, or if the mechanism is given all the queries in advance and produces a list of answers, or implicitly, which occurs if the mechanism produces a synopsis from which the analysts extracts $r(S)$. Note that $r(S)$ may depend on random choices made by the mechanism and the history of queries. Let $E(S, r(S))$ denote the *error*, also called *noise* or *distortion*, of the response $r(S)$, so $E(S, r(S)) = |A(S) - r(S)|$.

The question we want to ask is, "How much noise is needed in order to preserve privacy?" Differential privacy is a specific privacy guarantee, but one might also consider weaker notions, so rather than guaranteeing privacy the modest goal in the lower bound arguments will simply be to prevent privacy catastrophes.

Definition 8.1. A mechanism is *blatantly non-private* if an adversary can construct a candidate database c that agrees with the real database d in all but $o(n)$ entries, i.e., $\|c - d\|_0 \in o(n)$.

In other words, a mechanism is blatantly non-private if it permits a reconstruction attack that allows the adversary to correctly guess the secret bit of all but $o(n)$ members of the database. (There is no requirement that the adversary know on which answers it is correct.)

Theorem 8.1. Let $\mathcal{M}$ be a mechanism with distortion of magnitude bounded by E. Then there exists an adversary that can reconstruct the database to within $4E$ positions.

An easy consequence of the theorem is that a privacy mechanism adding noise with magnitude always bounded by, say, $n/401$, permits an adversary to correctly reconstruct 99% of the entries.

Proof. Let d be the true database. The adversary attacks in two phases:

1. **Estimate the number of 1s in all possible sets:** Query $\mathcal{M}$ on all subsets $S \subseteq [n]$.
2. **Rule out "distant" databases:** For every candidate database $c \in \{0,1\}^n$, if $\exists S \subseteq [n]$ such that $|\sum_{i \in S} c_i - \mathcal{M}(S)| > E$, then rule out c. If c is not ruled out, then output c and halt.

Since $\mathcal{M}(S)$ never errs by more than E, the real database will not be ruled out, so this simple (but inefficient!) algorithm will output *some* candidate database c. We will argue that the number of positions in which c and d differ is at most $4 \cdot E$.

Let I_0 be the indices in which $d_i = 0$, that is, $I_0 = \{i \mid d_i = 0\}$. Similarly, define $I_1 = \{i \mid d_i = 1\}$. Since c was not ruled out, $|\mathcal{M}(I_0) - \sum_{i \in I_0} c_i| \leq E$. However, by assumption $|\mathcal{M}(I_0) - \sum_{i \in I_0} d_i| \leq E$. It follows from the triangle inequality that c and d differ in at most $2E$ positions in I_0; the same argument shows that they differ in at most $2E$ positions in I_1. Thus, c and d agree on all but at most $4E$ positions. □

What if we consider more realistic bounds on the number of queries? We think of $\sqrt{n}$ as an interesting threshold on noise, for the following reason: if the database contains n people drawn uniformly at random from a population of size $N \gg n$, and the fraction of the population satisfying a given condition is p, then we expect the number of rows in the database satisfying the property to be roughly $np \pm \Theta(\sqrt{n})$, by the properties of the binomial distribution. That is, the sampling error is on the order of $\sqrt{n}$. We would like that the noise introduced for privacy is smaller than the sampling error, ideally $o(\sqrt{n})$. The next result investigates the feasibility of such small error when the number of queries is linear in n. The result is negative.

Ignoring computational complexity, to see why there might exist a query-efficient attack we modify the problem slightly, looking at databases $d \in \{-1,1\}^n$ and query vectors $v \in \{-1,1\}^n$. The true answer is again defined to be $d \cdot v$, and the response is a noisy version of the true answer. Now, consider a candidate database c that is far from d, say, $\|c-d\|_0 \in \Omega(n)$. For a random $v \in_R \{-1,1\}^n$, with constant probability we have $(c-d)\cdot v \in \Omega(\sqrt{n})$. To see this, fix $x \in \{-1,1\}^n$ and choose $v \in_R \{-1,1\}^n$. Then $x \cdot v$ is a sum of independent random variables $x_i v_i \in_R \{-1,1\}$, which has expectation 0 and variance n, and is distributed according to a scaled and shifted binomial distribuiton. For the same reason, if c and d differ in at least αn rows, and v is chosen at random, then $(c-d)\cdot v$ is binomially distributed with mean 0 and variance at least αn. Thus, we expect $c \cdot v$ and $d \cdot v$ to differ by at least $\alpha\sqrt{n}$ with constant probability, by the properties of the binomial distribution. Note that we are using the *anti*-concentration property of the distribution, rather than the usual appeal to concentration.

This opens an attack for ruling out c when the noise is constrained to be $o(\sqrt{n})$: compute the difference between $c \cdot v$ and the noisy response $r(v)$. If the magnitude of this difference exceeds $\sqrt{n}$ — which will occur with constant probability over the choice of v — then rule out c. The next theorem formalizes this argument and further shows that the attack is resilient even to a large fraction of completely arbitrary responses: Using a linear number of ±1 questions, an attacker can reconstruct almost the whole database if the curator is constrained to answer at least $\frac{1}{2}+\eta$ of the questions within an absolute error of $o(\sqrt{n})$.

Theorem 8.2. For any $\eta > 0$ and any function $\alpha = \alpha(n)$, there is constant b and an attack using bn ±1 questions that reconstructs a database that agrees with the real database in all but at most $(\frac{2\alpha}{\eta})^2$ entries, if the curator answers at least $\frac{1}{2} + \eta$ of the questions within an absolute error of α.

Proof. We begin with a simple lemma.

Lemma 8.3. Let $Y = \sum_{i=1}^{k} X_i$ where each X_i is a ± 2 independent Bernoulli random variable with mean zero. Then for any y and any $\ell \in \mathbb{N}$, $Pr[Y \in [2y, 2(y+\ell)]] \leq \frac{\ell+1}{\sqrt{k}}$.

Proof. Note that Y is always even and that $Pr[Y = 2y] = \binom{k}{(k+y)/2}(\frac{1}{2})^k$. This expression is at most $\binom{k}{\lceil k/2 \rceil}(\frac{1}{2})^k$. Using Stirling's approximation, which says that $n!$ can be approximated by $\sqrt{2n\pi}(n/e)^n$, this is bounded by $\sqrt{\frac{2}{\pi k}}$. The claim follows by a union bound over the $\ell + 1$ possible values for Y in $[2y, 2(y+\ell)]$. □

The adversary's attack is to choose bn random vectors $v \in \{-1, 1\}^n$, obtain responses $(y_1, \ldots, y_{bn})$, and then output any database c such that $|y_i - (Ac)_i| \leq \alpha$ for at least $\frac{1}{2} + \eta$ of the indices i, where A is the $bn \times n$ matrix whose rows are the random query vectors v.

Let the true database be d and let c be the reconstructed database. By assumption on the behavior of the mechanism, $|(Ad)_i - y_i| \leq \alpha$ for a $1/2 + \eta$ fraction of $i \in [bn]$. Since c was not ruled out, we also have that $|(Ac)_i - y_i| \leq \alpha$ for a $1/2 + \eta$ fraction of $i \in [bn]$. Since any two such sets of indices agree on at least a 2η fraction of $i \in [bn]$, we have from the triangle inequality that for at least $2\eta bn$ values of i, $|[(c-d)A]_i| \leq 2\alpha$.

We wish to argue that c agrees with d in all but $(\frac{2\alpha}{\eta})^2$ entries. We will show that if the reconstructed c is far from d, disagreeing on at least $(2\alpha/\eta)^2$ entries, the probability that a randomly chosen A will satisfy $|[A(c-d)]_i| \leq 2\alpha$ for at least $2\eta bn$ values of i will be extremely small — so small that, for a random A, it is extremely unlikely that there even exists a c far from d that is not eliminated by the queries in A.

Assume the vector $z = (c - d) \in \{-2, 0, 2\}^n$ has Hamming weight at least $(\frac{2\alpha}{\eta})^2$, so c is far from d. We have argued that, since c is produced by the attacker, $|(Az)_i| \leq 2\alpha$ for at least $2\eta bn$ values of i. We shall call such a z *bad with respect to A*. We will show that, with high probability over the choice of A, no z is bad with respect to A.

For any i, $v_i z$ is the sum of at least $(\frac{2\alpha}{\eta})^2$ ±2 random values. Letting $k = (2\alpha/\eta)^2$ and $\ell = 2\alpha$, we have by Lemma 8.3 that the probability that $v_i z$ lies in an interval of size 4α is at most η, so the expected number of queries for which $|v_i z| \leq 2\alpha$ is at most ηbn. Chernoff bounds now imply that the probability that this number exceeds $2\eta bn$ is at most $\exp(-\frac{\eta bn}{4})$. Thus the probability of a particular $z = c - d$ being bad with respect to A is at most $\exp(-\frac{\eta bn}{4})$.

Taking a union bound over the atmost 3^n possible zs, we get that with probability at least $1 - \exp(-n(\frac{\eta b}{4} - \ln 3))$, no bad z exists. Taking $b > 4 \ln 3/\eta$, the probability that such a bad z exists is exponentially small in n. □

Preventing blatant non-privacy is a very low bar for a privacy mechanism, so if differential privacy is meaningful then lower bounds for preventing blatant non-privacy will also apply to any mechanism ensuring differential privacy. Although for the most part we ignore computational issues in this monograph, there is also the question of the efficiency of the attack. Suppose we were able to prove that (perhaps under some computational assumption) there exist low-distortion mechanisms that are "hard" to break; for example, mechanisms for which producing a candidate database c close to the original database is hard? Then, although a low-distortion mechanism might fail to be differentially private in theory, it could conceivably provide privacy against bounded adversaries. Unfortunately, this is not the case. In particular, when the noise is always in $o(\sqrt{n})$, there is an efficient attack using exactly n fixed queries; moreover, there is even a computationally efficient attack requiring a linear number of queries in which a 0.239 fraction may be answered with wild noise.

In the case of "internet scale" data sets, obtaining responses to n queries is infeasible, as n is extremely large, say, $n \geq 10^8$. What happens if the curator permits only a sublinear number of questions? This inquiry led to the first algorithmic results in (what has evolved to be) (ε, δ)-differential privacy, in which it was shown how to maintain privacy against a sublinear number of counting queries by adding binomial noise of order $o(\sqrt{n})$ — less than the sampling error! — to each true answer. Using the tools of differential privacy we can do this either using either (1) the Gaussian mechanism or (2) the Laplace mechanism and advanced composition.

8.2 Lower bounds for differential privacy

The results of the previous section yielded lower bounds on distortion needed to ensure any reasonable notion of privacy. In contrast, the result in this section is specific to differential privacy. Although some of the details in the proof are quite technical, the main idea is elegant: suppose (somehow) the adversary has narrowed down the set of possible databases to a relatively small set S of 2^s vectors, where the L_1 distance between each pair of vectors is some large number Δ. Suppose further that we can find a k-dimensional query F, 1-Lipschitz in each of its output coordinates, with the property that the true answers to the query look very different (in L_∞ norm) on the different vectors in our set; for example, the distance on any two elements in the set may be $\Omega(k)$. It is helpful to think geometrically about the "answer space" $\mathbb{R}^k$. Each element x in the set S gives rise to a vector $F(x)$ in answer space. The actual response will be a perturbation of this point in answer space. Then a volume-based pigeon hole argument (in answer space) shows that, if with even moderate probability the (noisy) responses are "reasonably" close to the true answers, then ϵ cannot be very small.

This stems from the fact that for $(\varepsilon, 0)$-differentially private mechanisms $\mathcal{M}$, for *arbitrarily different* databases x, y, any response in the support of $\mathcal{M}(x)$ is also in the support of $\mathcal{M}(y)$. Taken together with the construction of an appropriate collection of vectors and a (contrived, non-counting) query, the result yields a lower bound on distortion that is linear k/ε. The argument appeals to Theorem 2.2, which discusses group privacy. In our case the group in question corresponds to the indices contributing to the (L_1) distance between a pair of vectors in S.

8.2.1 Lower bound by packing arguments

We begin with an observation which says, intuitively, that if the "likely" response regions, when the query is F, are disjoint, then we can bound

ϵ from below, showing that privacy can't be too good. When $\|F(x_i) - F(x_j)\|_\infty$ is large, this says that to get very good privacy, even when restricted to databases that differ in many places, we must get very erroneous responses on some coordinate of F.

The argument uses the histogram representation of databases. In the sequel, $d = |\mathcal{X}|$ denotes the size of the universe from which database elements are drawn.

Lemma 8.4. Assume the existence of a set $S = \{x_1, \ldots, x_{2^s}\}$, where each $x_i \in \mathbb{N}^d$, such that for $i \neq j$, $\|x_i - x_j\|_1 \leq \Delta$. Further, let $F : \mathbb{N}^d \to \mathbb{R}^k$ be a k-dimensional query. For $1 \leq i \leq 2^s$, let B_i denote a region in $\mathbb{R}^k$, the answer space, and assume that the B_i are mutually disjoint. If $\mathcal{M}$ is an $(\varepsilon, 0)$-differentially private mechanism for F such that, $\forall 1 \leq i \leq 2^s$, $\Pr[\mathcal{M}(x_i) \in B_i] \geq 1/2$, then $\varepsilon \geq \frac{\ln(2)(s-1)}{\Delta}$.

Proof. By assumption $\Pr[\mathcal{M}(x_j) \in B_j] \geq 2^{-1}$. Since the regions $B_1, \ldots, B_{2^s}$ are disjoint, $\exists j \neq i \in [2^s]$ such that $\Pr[\mathcal{M}(x_i) \in B_j] \leq 2^{-s}$. That is, for at least one of the $2^s - 1$ regions B_j, the probability that $\mathcal{M}(x_i)$ is mapped to this B_j is at most 2^{-s}. Combining this with differential privacy, we have

$$\frac{2^{-1}}{2^{-s}} \leq \frac{\Pr_{\mathcal{M}}[B_j|x_j]}{\Pr_{\mathcal{M}}[B_j|x_i]} \leq \exp(\varepsilon\Delta). \qquad \square$$

Corollary 8.5. Let $S = \{x_1, \ldots, x_{2^s}\}$ be as in Lemma 8.4, and assume that for any $i \neq j$, $\|F(x_i) - F(x_j)\|_\infty \geq \eta$. Let B_i denote the L_∞ ball in $\mathbb{R}^k$ of radius $\eta/2$ centered at x_i. Let $\mathcal{M}$ be any ε-differentially private mechansim for F satisfying

$$\forall 1 \leq i \leq 2^s : \Pr[\mathcal{M}(x_i) \in B_i] \geq 1/2.$$

Then $\varepsilon \geq \frac{(\ln 2)(s-1)}{\Delta}$.

Proof. The regions $B_1, \ldots, B_{2^s}$ are disjoint, so the conditions of Lemma 8.4 are satisfied. The corollary follows by applying the lemma and taking logarithms. $\square$

In Theorem 8.8 below we will look at queries F that are simply k independently and randomly generated (nonlinear!) queries. For suitable S and F (we will work to find these) the corollary says that if with probability at least $1/2$ *all* responses simultaneously have small error, then privacy can't be too good. In other words,

Claim 8.6 (Informal Restatement of Corollary 8.5)**.** To obtain $(\varepsilon, 0)$-differential privacy for $\varepsilon \leq \frac{\ln(2)(s-1)}{\Delta}$, the mechanism must add noise with L_∞ norm greater than $\eta/2$ with probability exceeding $1/2$.

As a warm-up exercise, we prove an easier theorem that requires a large data universe.

Theorem 8.7. Let $\mathcal{X} = \{0,1\}^k$. Let $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}^k$ be an $(\varepsilon, 0)$-differentially private mechanism such that for every database $x \in \mathcal{X}^n$ with probability at least $1/2$ $\mathcal{M}(x)$ outputs all of the 1-way marginals of x with error smaller than $n/2$. That is, for each $j \in [k]$, the jth component of $\mathcal{M}(x)$ should approximately equal the number of rows of x whose jth bit is 1, up to an error smaller than $n/2$. Then $n \in \Omega(k/\varepsilon)$.

Note that this bound is tight to within a constant factor, by the simple composition theorem, and that it separates $(\varepsilon, 0)$-differential privacy from (ε, δ)-differential privacy, for $\delta \in 2^{-o(n)}$, since, by the advanced composition theorem (Theorem 3.20), Laplace noise with parameter $b = \sqrt{k \ln(1/\delta)}/\varepsilon$ suffices for the former, in contrast to $\Omega(k/\varepsilon)$ needed for the latter. Taking $k \in \Theta(n)$ and, say, $\delta = 2^{-\log^2 n}$, yields the separation.

Proof. For every string $w \in \{0,1\}^k$, consider the database x_w consisting of n identical rows, all of which equal w. Let $B_w \in \mathbb{R}^k$ consist of all tuples of numbers that provide answers to the 1-way marginals on x with error less than $n/2$. That is,

$$B_w = \{(a_1, \ldots, a_k)\} \in \mathbb{R}^k : \forall i \in [k]\, |a_i - nw_i| < n/2\}.$$

Put differently, B_w is the open ℓ_∞ of radius $n/2$ around $nw \in \{0,n\}^k$. Notice that the sets B_w are mutually disjoint.

If M is an accurate mechanism for answering 1-way marginals, then for every w the probability of landing in B_w when the database is x_w should be at least $1/2$: $\Pr[\mathcal{M}(x_w) \in B_w] \geq 1/2$. Thus, setting $\Delta = n$ and $s = k$ in Corollary 8.5 we have $\varepsilon \geq \frac{(\ln 2)(s-1)}{\Delta}$. □

Theorem 8.8. For any $k, d, n \in \mathbb{N}$ and $\varepsilon \in (0, 1/40]$, where $n \geq \min\{k/\varepsilon, d/\varepsilon\}$, there is a query $F : \mathbb{N}^d \to \mathbb{R}^k$ with per-coordinate sensitivity at most 1 such that any $(\varepsilon, 0)$-differentially private mechanism adds noise of L_∞ norm $\Omega(\min\{k/\varepsilon, d/\varepsilon\})$ with probability at least 1/2 on some databases of weight at most n.

Note that $d = |\mathcal{X}|$ need not be large here, in contrast to the requirement in Theorem 8.7.

Proof. Let $\ell = \min\{k, d\}$. Using error-correcting codes we can construct a set $S = \{x_1, \ldots, x_{2^s}\}$, where $s = \ell/400$, such that each $x_i \in \mathbb{N}^d$ and in addition

1. $\forall i : \|x_i\|_1 \leq w = \ell/(1280\varepsilon)$
2. $\forall i \neq j,\ \|x_i - x_j\|_1 \geq w/10$

We do not give details here, but we note that the databases in S are of size at most $w < n$, and so $\|x_i - x_j\|_1 \leq 2w$. Taking $\Delta = 2w$ the set S satisfies the conditions of Corollary 8.5. The remainder of our effort is to obtain the queries F to which we will apply Corollary 8.5. Given $S = \{x_1, \ldots, x_{2^s}\}$, where each $x_i \in \mathbb{N}^d$, the first step is to define a mapping from the space of histograms to vectors in $\mathbb{R}^{2^s}$, $\mathcal{L}_S : \mathbb{N}^d \to \mathbb{R}^{2^s}$. Intuitively (and imprecisely!), given a histogram x, the mapping lists, for each $x_i \in S$, the L_1 distance from x to x_i. More precisely, letting w be an upper bound on the weight of any x_i in our collection we define the mapping as follows.

- For every $x_i \in S$, there is a coordinate i in the mapping.
- The ith coordinate of $\mathcal{L}_S(x)$ is $\max\{w/30 - \|x_i - z\|_1, 0\}$.

Claim 8.9. If $x_1, \ldots, x_{2^s}$ satisfy the conditions

1. $\forall i \|x_i\|_1 \leq w$; and
2. $\forall i \neq j \|x_i - x_j\|_1 \geq w/10$

then the map $\mathcal{L}_S$ is 1-Lipschitz; in particular, if $\|z_1 - z_2\|_1 = 1$, then $\|\mathcal{L}_S(z_1) - \mathcal{L}_S(z_2)\|_1 \leq 1$, assuming $w \geq 31$.

Proof. Since we assume $w \geq 31$ we have that if $z \in \mathbb{N}^d$ is close to some $x_i \in S$, meaning $w/30 > \|x_i - z\|_1$, then z cannot be close to any other $x_j \in S$, and the same is true for all $\|z' - z\|_1 \leq 1$. Thus, for any z_1, z_2 such that $\|z_1 - z_2\| \leq 1$, if A denotes the set of coordinates where at least one of $\mathcal{L}_S(z_1)$ or $\mathcal{L}_S(z_2)$ is non-zero, then A is either empty or is a singleton set. Given this, the statement in the claim is immediate from the fact that the mapping corresponding to any particular coordinate is clearly 1-Lipschitz. □

We can finally describe the queries F. Corresponding to any $r \in \{-1,1\}^{2^s}$, we define $f_r : \mathbb{N}^d \to \mathbb{R}$, as

$$f_r(x) = \sum_{i=1}^{d} \mathcal{L}_S(x)_i \cdot r_i\,,$$

which is simply the inner product $\mathcal{L}_S \cdot r$. F will be a random map $F : \mathbb{N}^d \to \mathbb{R}^k$: Pick $r_1, \ldots, r_k \in \{-1,1\}^{2^s}$ independently and uniformly at random and define

$$F(x) = (f_{r_1}(x), \ldots, f_{r_k}(x))\,.$$

That is, $F(x)$ is simply the result of the inner product of $\mathcal{L}_S(x)$ with k randomly chosen ± 1 vectors.

Note that for any $x \in S$ $\mathcal{L}_S(x)$ has one coordinate with value $w/30$ (and the others are all zero), so $\forall r_i \in \{-1,1\}^{2^s}$ and $x \in S$ we have $|f_{r_i}(x)| = w/30$. Now consider any $x_h, x_j \in S$, where $h \neq j$. It follows that for any $r_i \in \{-1,1\}^{2^s}$,

$$\Pr_{r_i}[|f_{r_i}(x_h) - f_{r_i}(x_j)| \geq w/15] \geq 1/2$$

(this event occurs when $(r_i)_h = -(r_i)_j$). A basic application of the Chernoff bound implies that

$$\begin{aligned}&\Pr_{r_1,\ldots,r_k}[\text{For at least } 1/10 \text{ of the } r_i\text{s},\\ &|f_{r_i}(x_h) - f_{r_i}(x_j)| \geq w/15] \geq 1 - 2^{-k/30}\,.\end{aligned}$$

Now, the total number of pairs (x_i, x_j) of databases such that $x_i, x_j \in S$ is at most $2^{2s} \leq 2^{k/200}$. Taking a union bound this implies

$$\begin{aligned}&\Pr_{r_1,\ldots,r_k}[\forall h \neq j, \quad \text{For at least } 1/10 \text{ of the } r_i\text{s},\\ &|f_{r_i}(x_h) - f_{r_i}(x_j)| \geq w/15] \geq 1 - 2^{-k/40}\end{aligned}$$

This implies that we can fix $r_1, \ldots, r_k$ such that the following is true.

$$\forall h \neq j, \quad \text{For at least } 1/10 \text{ of the } r_i\text{s}, \quad |f_{r_i}(x_h) - f_{r_i}(x_j)| \geq w/15$$

Thus, for any $x_h \neq x_j \in S$, $\|F(x_h) - F(x_j)\|_\infty \geq w/15$.

Setting $\Delta = 2w$ and $s = \ell/400 > 3\varepsilon w$ (as we did above), and $\eta = w/15$, we satisfy the conditions of Corollary 8.5 and conclude $\Delta \leq (s-1)/\varepsilon$, proving the theorem (via Claim 8.6). □

The theorem is almost tight: if $k \leq d$ then we can apply the Laplace mechanism to each of the k sensitivity 1 component queries in F with parameter k/ε, and we expect the maximum distortion to be $\Theta(k \ln k/\varepsilon)$. On the other hand, if $d \leq k$ then we can apply the Laplace mechanism to the d-dimensional histogram representing the database, and we expect the maximum distortion to be $\Theta(d \ln d/\varepsilon)$.

The theorem actually shows that, given knowledge of the set S and knowledge that the actual database is an element $x \in S$, the adversary can completely determine x if the L_∞ norm of the distortion is too small. How in real life might the adversary obtain a set S of the type used in the attack? This can occur when a *non-private* database system has been running on a dataset, say, x. For example, x could be a vector in $\{0,1\}^n$ and the adversary may have learned, through a sequence of linear queries, that $x \in \mathcal{C}$, a linear code of distance, say $n^{2/3}$. Of course, if the database system is not promising privacy there is no problem. The problem arises if the administrator decides to replace the existing system with a differentially private mechanism — after several queries have received noise-free responses. In particular, if the administrator chooses to use (ε, δ)-differential privacy for subsequent k queries then the distortion might fall below the $\Omega(k/\varepsilon)$ lower bound, permitting the attack described in the proof of Theorem 8.8.

The theorem also emphasizes that there is a fundamental difference between auxiliary information about (sets of) members of the database and information about the database *as a whole*. Of course, we already knew this: being told that the number of secret bits sums to exactly $5,000$ completely destroys differential privacy, and an adversary that already knew the secret bit of every member of the database except one individual could then conclude the secret bit of the remaining individual.

Additional Consequences. Suppose $k \leq d$, so $\ell = k$ in Theorem 8.8. The linear in k/ε lower bound on noise for k queries sketched in the previous section immediately yields a separation between counting queries and arbitrary 1-sensitivity queries, as the SmallDB construction answers (more than) n queries with noise roughly $n^{2/3}$ while maintaining differential privacy. Indeed, this result also permits us to conclude that there is no small α-net for large sets of arbitrary low sensitivity queries, for $\alpha \in o(n)$ (as otherwise the net mechanism would yield an $(\varepsilon, 0)$ algorithm of desired accuracy).

8.3 Bibliographic notes

The first reconstruction attacks, including Theorem 8.1, are due to Dinur and Nissim [18], who also gave an attack requiring only polynomial time computation and $O(n \log^2 n)$ queries, provided the noise is always $o(\sqrt{n})$. Realizing that attacks requiring n random linear queries, when n is "internet scale," are infeasible, Dinur, Dwork, and Nissim gave the first positive results, showing that for a sublinear number of subset sum queries, a form of privacy (now known to imply (ε, δ)-differential privacy) can be achieved by adding noise scaled to $o(\sqrt{n})$ [18]. This was exciting because it suggested that, if we think of the database as drawn from an underlying population, then, even for a relatively large number of counting queries, privacy could be achieved with distortion smaller than the sampling error. This eventulaly lead, via more general queries [31, 6], to differential privacy. The view of these queries as a privacy-preserving programming primitive [6] inspired McSherry's Privacy Integrated Queries programming platform [59].

The reconstruction attack of Theorem 8.2 appears in [24], where Dwork, McSherry, and Talwar showed that polynomial time reconstruction is possible even if a 0.239 fraction of the responses have wild, arbitrary, noise, provided the others have noise $o(\sqrt{n})$.

The geometric approach, and in particular Lemma 8.4, is due to Hardt and Talwar [45], who also gave a geometry-based algorithm proving these bounds tight for small numbers $k \leq n$ of queries, under a commonly believed conjecture. Dependence on the conjecture was later removed by Bhaskara et al. [5]. The geometric approach was extended to arbitrary numbers of queries by Nikolov et al. [66], who gave an algorithm with instance-optimal mean squared error. For the few queries case this leads, via a boosting argument, to low expected worst-case error. Theorem 8.8 is due to De [17].