

A Note on CDS for Random Functions

Marshall Ball

Courant Institute of Mathematical Sciences
New York University

Abstract

We record the following bound: almost all Boolean functions $f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}$ require $2n - 2 \log n - O(1)$ bits of communication for perfect two-party conditional disclosure of secrets with a one-bit secret. This only gives a factor 2 improvement over the previous best bound due to Applebaum, Holenstein, Mishra and Shayevitz¹ [EUROCRYPT 2018], but at least (nearly) matches the input length. The proof is a straightforward compression argument that uses the birthday bound to give a succinct representation of transcript collisions (the latter being a technique from Feige, Kilian and Naor [STOC 1994], refined in Applebaum et al.).

Viewing transcript-collision testing as an elementary case of distribution testing extends the argument to imperfect correctness and privacy. We do not currently know how to extend the argument to an explicit hard predicate.

1 Introduction

We recall the definition of conditional disclosure of secrets (CDS) introduced by Gertner, Ishai, Kushilevitz and Malkin [GIKM00].

Conditional disclosure of secrets. Let $f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}$ be a Boolean function. In a *conditional disclosure of secrets* (CDS) protocol, Alice holds an input $x \in \{0, 1\}^n$ and Bob holds an input $y \in \{0, 1\}^n$, and neither knows the other's input. Both hold a secret bit $S \in \{0, 1\}$ and shared randomness R . Each sends a single message to a receiver who knows (x, y) but not S or R . The receiver should learn S if $f(x, y) = 1$, and should learn nothing about S otherwise.

We fix the model once and for all. The shared randomness R is independent of the inputs and secret and has finite support; Alice's and Bob's messages are

$$A(s, x, R) \in \{0, 1\}^a, \quad B(s, y, R) \in \{0, 1\}^b,$$

of *fixed* lengths a and b , so the transcript $(A(s, x, R), B(s, y, R))$ ranges over an alphabet of size at most 2^{a+b} . For an input (x, y) and a secret $s \in \{0, 1\}$, write

$$D_{x,y}^s := (A(s, x, R), B(s, y, R))$$

for the transcript distribution. We use total variation distance, $\Delta(P, Q) := \frac{1}{2} \sum_{\tau} |P(\tau) - Q(\tau)|$.

¹Unlike the present note, Applebaum et al. also show a matching bound for an explicit function: inner product. They also prove bounds on Private Simultaneous Message protocols.

Definition 1. Let $\eta, \delta \in [0, 1]$. A pair of maps $A : \{0, 1\} \times \{0, 1\}^n \times \text{Supp}(R) \rightarrow \{0, 1\}^a$ and $B : \{0, 1\} \times \{0, 1\}^n \times \text{Supp}(R) \rightarrow \{0, 1\}^b$ is an (η, δ) -CDS for f if:

1. (δ -correctness) for every (x, y) with $f(x, y) = 1$, there is a possibly randomized decoder $\text{Dec}_{x,y}$ such that

$$\Pr_{T \sim D_{x,y}^s} [\text{Dec}_{x,y}(T) = s] \geq 1 - \delta \quad \text{for each } s \in \{0, 1\}.$$

The probability includes the decoder's internal randomness, independent of the transcript. The same decoder must work for both secrets.

2. (η -privacy) for every (x, y) with $f(x, y) = 0$,

$$D_{x,y}^0 \approx_\eta D_{x,y}^1.$$

Here $P \approx_\eta Q$ means $\Delta(P, Q) \leq \eta$.

The communication of the protocol is $a + b$, and $C_{\text{CDS}}^{\eta, \delta}(f)$ is the least communication of an (η, δ) -CDS for f . A protocol is perfect if $\eta = \delta = 0$; we abbreviate $C_{\text{CDS}}^{0,0}(f)$ to $C_{\text{CDS}}(f)$.

The perfect case. Perfect correctness and privacy have a particularly clean description in terms of transcript supports and distributions:

1. *Perfect correctness* is equivalent to disjoint supports on every 1-input:

$$f(x, y) = 1 \implies \text{Supp}(D_{x,y}^0) \cap \text{Supp}(D_{x,y}^1) = \emptyset.$$

2. *Perfect privacy* is equivalent to identical distributions on every 0-input:

$$f(x, y) = 0 \implies D_{x,y}^0 \equiv D_{x,y}^1.$$

Lower Bound for Perfect CDS. We begin by explaining the result for perfect CDS.

Informal Theorem 1 (Theorem 1). *Almost all functions require perfect CDS communication at least $2n - 2 \log n - O(1)$.*

The proof is a compression argument using the transcript-collision viewpoint of Feige, Kilian and Naor [FKN94]. Applebaum, Holenstein, Mishra and Shayevitz revisited and repaired this viewpoint [AHMS20].

Fix an input x, y , and consider a transcript sampled under secret 0 and an independent transcript sampled under secret 1. Consider the event that these sampled transcripts collide:

$$(A(0, x, R), B(0, y, R)) = (A(1, x, R'), B(1, y, R'))$$

1. If $f(x, y) = 1$, perfect correctness (i.e. disjointness of supports) tells us this collision never occurs.
2. If $f(x, y) = 0$, perfect privacy tells us that the two distributions are identical. Their common support has size at most 2^c , where $c = a + b$ is the communication. The birthday bound therefore gives a collision with probability at least $1/4$ when we take $t = \sqrt{2^c}$ samples under each secret.

We then use this simple observation to compress the description of f . We sample these $2t$ tapes once and evaluate both local message functions on every local input. The resulting table has $t \cdot 2^n$ rows,

$$(A(0, z, r_i^0), B(0, z, r_i^0), A(1, z, r_i^1), B(1, z, r_i^1))_{i \in [t], z \in \{0,1\}^n},$$

each of length $2c$. Crucially, this uses $2^{n+1}tc$ bits to describe samples of the transcript at all 2^{2n} input pairs. The table determines a set $C(T)$ of input pairs where a collision occurs:

$$\exists i, j : (A(0, x, r_i^0), B(0, y, r_i^0)) = (A(1, x, r_j^1), B(1, y, r_j^1)).$$

1. Correctness gives that $C(T)$ can only contain x, y such that $f(x, y) = 0$ (i.e. never includes x, y s.t. $f(x, y) = 1$, or $C(T) \subseteq f^{-1}(0)$).
2. The birthday bound gives, by linearity of expectation, some table with $|C(T)| \geq |f^{-1}(0)|/4$.

In other words, there exists a table T (of length $2^{n+1}tc$) that encodes at least a quarter of the zeros of f . Hence, to describe f we only need to describe the remaining truth-table entries, one bit each.

A random function nearly always has at least $2^{2n}/3$ zeros. Let us assume this holds for the purposes of this high-level explanation. Its corresponding table from above would therefore encode at least $2^{2n}/12$ truth-table entries with only $2^{n+1}tc$ bits. As the locations of these bits are described by T , we can describe the remainder of the truth table with at most $\frac{11}{12}2^{2n}$ bits. So, if

$$2^{n+1}tc \leq \frac{2^{2n}}{24},$$

the description saves at least $2^{2n}/24$ bits, apart from the $O(\log(c+1))$ bits needed to specify the message split. Counting descriptions shows that only an exponentially small fraction of functions can admit such a saving.

Thus, ignoring rounding in $t = 2^{c/2}$, almost all functions must satisfy

$$2^{n+1}2^{c/2}c > \frac{2^{2n}}{24} \implies 2^{c/2}c > \frac{2^n}{48} \implies c \geq 2n - O(\log n).$$

Section 3 gives the full proof; Appendix A contains a proof of this precise birthday bound for completeness.

Imperfect case: δ -correctness and η -privacy. The collision test above distinguishes identical transcript distributions from distributions with disjoint supports. To allow error, we replace it with a distribution test that handles the corresponding relaxed promise. On 0-inputs, privacy makes the distributions η -close in total variation distance. On 1-inputs, correctness forces them to be $\gamma := 1 - 2\delta$ far. Section 4 derives this separation from Definition 1.

Let $q(M)$ be a *distribution-testing sample bound for these parameters*: using $q(M)$ independent samples from each of two unknown distributions on an alphabet of size M , the test distinguishes distance at most η from distance at least γ , with error at most $1/10$ on every promised pair. The table of local messages has $2^{n+1}q(2^c)c$ bits. The test may make mistakes, so we also encode its error pattern. Constant error below $1/2$ still saves a constant fraction of the truth table. Theorem 2 formalizes this conversion. In particular, a bound $q(M) = O(M^\alpha)$ yields communication at least $\frac{1}{\alpha}(n - \log n) - O(1)$, while $q(M) = O(M/\log M)$ yields $n - O(1)$. The case $\eta = 0$ is *closeness*

testing; allowing $\eta > 0$ gives *tolerant closeness testing*. We use the closeness-testing bound of Chan, Diakonikolas, Valiant and Valiant [CDVV14] and the tolerant-testing bound of Valiant and Valiant [VV11]. For background on distribution testing, see Canonne’s excellent surveys [Can20, Can22].

For fixed error parameters in the indicated ranges, we obtain the following bounds:

errors	samples per distribution	communication lower bound
$\eta = 0, \delta = 0$	$O(\sqrt{M})$	$2n - 2 \log n - O(1)$
$0 < \eta < 1, \delta = 0$	$O_\eta(\sqrt{M})$	$2n - 2 \log n - O_\eta(1)$
$\eta = 0, 0 < \delta < 1/2$	$O_\delta(M^{2/3})$	$\frac{3}{2}(n - \log n) - O_\delta(1)$
$0 < \eta < 1 - 2\delta, 0 < \delta < 1/2$	$O_{\eta,\delta}(M/\log M)$	$n - O_{\eta,\delta}(1)$

The strict gap $\eta < 1 - 2\delta$ makes the two testing cases disjoint. In the last row, the factor $\log M = c$ saved by the tester cancels the factor c in the table size, which removes the $\log n$ loss. These remain bounds for random predicates.

Prior work. CDS was introduced by Gertner, Ishai, Kushilevitz and Malkin [GIKM00]. The first non-trivial lower bounds are due to Gay, Kerenidis and Wee [GKW15], who gave a perfect CDS for inner product of total cost $n + 1$, an $\Omega(\sqrt{n})$ bound for *linear* CDS, and, for general CDS, only $\Omega(\log n)$; they asked whether $\Omega(n)$ could be proved.

The transcript-collision viewpoint already appears in Feige, Kilian and Naor [FKN94]. As described by Applebaum et al. [AHMS20, Section 1.2], their PSM lower-bound argument turns on the set S_r of transcripts realizable under a random string r having small intersection with $S_{r'}$ for an independent r' ; a point of $S_r \cap S_{r'}$ is a transcript collision between two independent executions. Applebaum, Arkis, Raykov and Vasudevan [AARV21] transferred that bound to CDS by reduction — a CDS of cost t yields a 1-bit PSM of cost $t + 2n + 1$ — obtaining $0.99n$ for a non-explicit predicate.

Applebaum, Holenstein, Mishra and Shayevitz [AHMS20] then showed that the FKN argument has a real gap (it assumes every non-trivial collision comes from sibling inputs), repaired it for PSM, and showed that the *original* argument does hold against weakly private fully revealing PSM — a model that a CDS reduces to directly. This gives a clean criterion for CDS, and it is the source of the bounds we improve.

Theorem [AHMS20, Thm. 5, full version]. Let $h : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}$ be a predicate such that no $h(x, \cdot)$ is identically zero, and let M bound the size of the largest 0-monochromatic rectangle. Then every perfect CDS for h has communication at least

$$2 \log |h^{-1}(0)| - \log M - \log |\mathcal{X}| - \log |\mathcal{Y}| - 1.$$

Two instantiations: almost every predicate on n -bit inputs needs $n - 4 - o(1)$ [AHMS20, Cor. 2, full version], and inner product needs $n - 3 - o(1)$ [AHMS20, Cor. 3, full version]. The latter, together with an $n + 1$ upper bound, settles the perfect CDS complexity of inner product up to 4 bits. Regrettably, we do not know how to extend the techniques in the present paper to any explicit function.

For imperfect security the state of the art is due to Applebaum and Vasudevan [AV21], who give elegant connections between CDS and communication complexity. A perfectly correct CDS

with simulator privacy error at most 0.1 costs $\Omega(\text{coNP}^{\text{cc}}(f)) - O(\log n)$, and a perfectly private CDS with correctness error at most 0.1 costs $\Omega(\text{PP}^{\text{cc}}(f)) - O(\log n)$ [AV21, Theorems 3 and 7, conference version]. These give $\Omega(n)$ for not-equal and for inner product respectively. When *both* guarantees have error at most 0.1, using their simulator-privacy convention, they recall the general bound $\Omega(\log(R^{A \rightarrow B} + R^{B \rightarrow A}))$. They also obtain a bound in terms of AM and coAM communication complexity, which gives $n^{\Omega(1)}$ for most predicates when both errors are 0.1 in their simulator-privacy convention [AV21, Theorem 9 and Corollary 10, conference version]. This yields polynomial lower bounds for nonexplicit functions.

In the other direction, Liu, Vaikuntanathan and Wee [LVW17] show that every predicate admits a *perfect* CDS with communication $2^{O(\sqrt{\log N \log \log N})} = 2^{O(\sqrt{n \log n})}$, where $N = 2^n$. The gap between $2n$ and $2^{O(\sqrt{n \log n})}$ remains, sadly, enormous.

2 Preliminaries

For $n \in \mathbb{N}$ write $[n] := \{1, \dots, n\}$; all logarithms are base 2. Throughout,

$$N := 2^n,$$

and we identify a function $f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}$ with its truth table, an element of $\{0, 1\}^{N^2}$ under a fixed ordering of $\{0, 1\}^n \times \{0, 1\}^n$. We write H_2 for the binary entropy function and Δ for total variation distance, and we use the standard bound $\sum_{j \leq \rho N^2} \binom{N^2}{j} \leq 2^{H_2(\rho)N^2}$, valid for $0 \leq \rho \leq 1/2$.

A birthday bound.

Proposition 1 (Birthday bound). *Let t be a positive integer and let D be a distribution supported on a set Ω with $|\Omega| \leq t^2$. If $X_1, \dots, X_t, Y_1, \dots, Y_t$ are independent samples from D , then*

$$\Pr[\exists i, j : X_i = Y_j] \geq \frac{1}{4}.$$

The proof appears in Appendix A.

3 The Perfect Case

Theorem 1. *Let $c_0 := \lfloor 2n - 2 \log n - 16 \rfloor$. For all sufficiently large n ,*

$$\Pr_f[C_{\text{CDS}}(f) \leq c_0] \leq 2^{-\Omega(2^{2n})}.$$

Here f is chosen uniformly at random from all functions $\{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}$.

Proof. Fix an integer $c \geq 0$ and any function $f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}$ that has a perfect CDS protocol of communication at most c . By padding, let A, B have fixed message lengths a, b with $a + b = c$. Let $t := \lceil \sqrt{2^c} \rceil$. For each $s \in \{0, 1\}$, sample t random tapes from the distribution of R , with all $2t$ tapes independent:

$$r_1^0, \dots, r_t^0 \quad r_1^1, \dots, r_t^1.$$

Iterating over all local inputs, these induce a table T with $t \cdot 2^n$ rows of length $2c$. Its row indexed by $(i, z) \in [t] \times \{0, 1\}^n$ is

$$(A(0, z, r_i^0), B(0, z, r_i^0), A(1, z, r_i^1), B(1, z, r_i^1)).$$

The same symbol z indexes Alice's and Bob's local-input lists; it does not restrict the protocol to inputs with $x = y$. At an arbitrary pair (x, y) , we read Alice's entry at x and Bob's entry at y .

Now for any such table we can define the following set of collisions:

$$C(T) := \left\{ (x, y) : \begin{array}{l} \text{there exist } i, j \in [t] \text{ such that} \\ A(0, x, r_i^0) = A(1, x, r_j^1) \text{ and} \\ B(0, y, r_i^0) = B(1, y, r_j^1) \end{array} \right\}.$$

Given the split $a + b = c$, this set is determined by the bits of T alone; the tapes and the protocol need not be included in its description. By perfect correctness, if $f(x, y) = 1$, then for every $r, r' \in \text{Supp}(R)$,

$$(A(0, x, r), B(0, y, r)) \neq (A(1, x, r'), B(1, y, r')).$$

Thus,

$$f(x, y) = 1 \implies (x, y) \notin C(T), \quad \text{or equivalently} \quad C(T) \subseteq f^{-1}(0). \quad (1)$$

This holds for every sampled table and is where we use the 1-inputs.

On the other hand, by perfect privacy, for any x, y such that $f(x, y) = 0$,

$$(A(0, x, R), B(0, y, R)) \equiv (A(1, x, R), B(1, y, R)),$$

and this distribution has support size at most $2^c \leq t^2$. Hence, by the birthday bound (Proposition 1),

$$\Pr_T[(x, y) \in C(T)] \geq \frac{1}{4}.$$

By linearity of expectation,

$$\mathbb{E}_T[|C(T)|] \geq \frac{|\{(x, y) : f(x, y) = 0\}|}{4}.$$

So there must exist a fixed choice of $r_1^0, \dots, r_t^0, r_1^1, \dots, r_t^1$ such that

$$C(T) \subseteq f^{-1}(0) \quad \text{and} \quad |C(T)| \geq \frac{|\{(x, y) : f(x, y) = 0\}|}{4}. \quad (2)$$

To reiterate, for every function f with a cost-at-most- c protocol, there exists a table satisfying (2). The table can depend on f . Collisions at different inputs need not be independent; only linearity of expectation was used.

Now consider a uniformly random function f . The expected size of $f^{-1}(0)$ is $2^{2n}/2 = N^2/2$. By a Chernoff bound,

$$\Pr_f[|\{(x, y) : f(x, y) = 0\}| < N^2/3] \leq \exp(-N^2/36).$$

Hence, except for an $e^{-N^2/36}$ fraction of functions, any cost-at-most- c CDS supplies a table T such that

$$|C(T)| \geq N^2/12 \quad \text{and} \quad C(T) \subseteq f^{-1}(0).$$

We now count all possible tables, including those that do not arise from a protocol.

For any choice of Alice's and Bob's message lengths a, b with $a + b = c$, a table has $t \cdot 2^n$ rows of length $2c$ and can therefore be specified in

$$2^{n+1}tc = 2Ntc$$

bits. There are $c + 1$ possible splits, so the total number of tables, including their splits, is at most

$$(c + 1)2^{2Ntc}.$$

Each fixed table determines a fixed collision set. If $|C(T)| \geq N^2/12$, then

$$\Pr_f[C(T) \subseteq f^{-1}(0)] = 2^{-|C(T)|} \leq 2^{-N^2/12},$$

because a uniformly random truth table has independent unbiased entries. Taking a union bound over all fixed tables handles the dependence of the witnessing table on f . Therefore,

$$\Pr_f[C_{\text{CDS}}(f) \leq c] \leq e^{-N^2/36} + (c + 1)2^{2Ntc - N^2/12}. \quad (3)$$

Take $c = c_0 = \lfloor 2n - 2 \log n - 16 \rfloor$, which is nonnegative for all sufficiently large n . By our choice of t ,

$$t \leq 2^{c/2} + 1 \leq \frac{N}{256n} + 1.$$

Since $c \leq 2n$, for all sufficiently large n we have

$$2Ntc \leq \frac{N^2}{64} + 4nN \leq \frac{N^2}{24}.$$

Substituting in (3) gives

$$\Pr_f[C_{\text{CDS}}(f) \leq c_0] \leq e^{-N^2/36} + (c_0 + 1)2^{-N^2/24} = 2^{-\Omega(N^2)}. \quad \square$$

4 Imperfect Correctness and Privacy

The proof of Theorem 1 used the table of samples to distinguish identical distributions from distributions with disjoint supports via collision testing. Noticing that this is a special case of a more general paradigm, we can replace the collision tester with more robust distribution testers to handle imperfect CDS.

4.1 From correctness and privacy to a testing promise

We use the (η, δ) -CDS model and transcript distributions $D_{x,y}^s$ of Definition 1.

Privacy gives $\Delta(D_{x,y}^0, D_{x,y}^1) \leq \eta$ at every 0-input. We next derive the separation forced by correctness at a 1-input.

Correctness gives $\Delta(D_{x,y}^0, D_{x,y}^1) \geq 1 - 2\delta$ at every 1-input. To see this, apply the same decoder to both distributions and use data processing:

$$\Delta(D_{x,y}^0, D_{x,y}^1) \geq \Pr_{D_{x,y}^0}[\text{Dec}_{x,y} = 0] - \Pr_{D_{x,y}^1}[\text{Dec}_{x,y} = 0] \geq (1 - \delta) - \delta.$$

Thus, writing $\gamma := 1 - 2\delta$, a protocol presents the following promise problem at each input: the pair $(D_{x,y}^0, D_{x,y}^1)$ is η -close if $f(x, y) = 0$ and γ -far if $f(x, y) = 1$. Deciding $f(x, y)$ from samples is a distribution-testing problem, namely *closeness (or equivalence) testing* when $\eta = 0$ and *tolerant closeness testing* when $\eta > 0$.

4.2 Distribution testing

Definition 2. Let $0 \leq \eta < \gamma \leq 1$, let $M \in \mathbb{N}$ and let $0 < \rho < 1/2$. A $(\eta, \gamma; M)$ -test with q samples and error ρ is a family of deterministic functions $\text{Test} : \Omega^q \times \Omega^q \rightarrow \{0, 1\}$, one for each finite set Ω with $|\Omega| \leq M$, such that for all distributions P, Q on Ω , with $X \sim P^{\otimes q}$ and $Y \sim Q^{\otimes q}$ independent,

$$\Delta(P, Q) \leq \eta \implies \Pr[\text{Test}(X, Y) = 1] \leq \rho, \quad \Delta(P, Q) \geq \gamma \implies \Pr[\text{Test}(X, Y) = 0] \leq \rho.$$

We write $q_\rho(\eta, \gamma; M)$ for the least such q : the distribution-testing sample complexity for proximity threshold η , separation threshold γ , alphabet size at most M , and error at most ρ .

Results in the distribution-testing literature are usually stated for randomized procedures with error $1/3$. Standard amplification and derandomization yield the form in our definition.

Lemma 1 (Amplification and derandomization). Suppose a randomized procedure using q samples from each distribution decides the $(\eta, \gamma; M)$ promise problem with error at most $1/3$ on every valid instance. Then $q_{1/10}(\eta, \gamma; M) = O(q)$.

Sketch. Run $k = O(1)$ independent copies and take the majority answer to push the error below $1/20$. Then, fix the samples and let Test output the majority answer over the internal coins, which is a deterministic function of the samples. If Test errs on a given sample tuple, the conditional error of the randomized procedure on that tuple is at least $1/2$; by Markov's inequality the probability of such a tuple is at most twice the overall error, hence at most $1/10$. $\square$

Lemma 2 (One-sided amplification). Let $0 < \beta < 1$ and $0 < \rho < 1/2$. Suppose a test never errs on γ -far instances and errs with probability at most $1 - \beta$ on η -close instances, using q samples from each distribution. Then $q_\rho(\eta, \gamma; M) \leq \lceil \log_{1/(1-\beta)}(1/\rho) \rceil \cdot q$.

Proof. Run k independent blocks and answer “close” if any block does. No block ever fires on a far instance, and all k fail to fire on a close instance with probability at most $(1 - \beta)^k$. $\square$

Bounds from the distribution-testing literature. We state the two results explicitly. In both, the distributions are unknown, their common alphabet is known, and the only access to them is through independent samples.

Theorem (Closeness testing [CDVV14, Theorem 1.1]). For every integer $M \geq 2$ and $0 < \varepsilon < 1$, there is a randomized algorithm that, given samples from arbitrary distributions P, Q on a common alphabet of size at most M , distinguishes

$$P = Q \quad \text{from} \quad \|P - Q\|_1 \geq \varepsilon.$$

It succeeds with probability at least $2/3$ on every promised pair and uses

$$O\left(\max\left\{\frac{M^{2/3}}{\varepsilon^{4/3}}, \frac{\sqrt{M}}{\varepsilon^2}\right\}\right)$$

independent samples from each distribution. The probability is over both the samples and the algorithm's internal randomness.

Taking $\varepsilon = \gamma/2$ puts the threshold in $(0, 1)$ and below the L_1 separation 2γ promised by a total variation distance of at least γ . By Lemma 1, the corresponding distribution-testing bound is

$$q_{1/10}(0, \gamma; M) = O\left(\max\left\{\frac{M^{2/3}}{\gamma^{4/3}}, \frac{\sqrt{M}}{\gamma^2}\right\}\right), \quad 0 < \gamma \leq 1. \quad (4)$$

The next result is the tolerant-testing consequence of Valiant and Valiant's L_1 -distance estimator [VV11, Theorem 4, Appendix C of the full version].

Theorem (Tolerant closeness testing, Valiant–Valiant). *Fix $0 \leq \eta < \gamma \leq 1$. For all sufficiently large M , there is a randomized algorithm that, given samples from arbitrary distributions P, Q on a common alphabet of size at most M , distinguishes*

$$\Delta(P, Q) \leq \eta \quad \text{from} \quad \Delta(P, Q) \geq \gamma.$$

It succeeds with probability at least $2/3$ on every promised pair and uses $O_{\gamma-\eta}(M/\log M)$ independent samples from each distribution. The probability is over both the samples and the algorithm's internal randomness. The implicit constant and the lower threshold on M may depend on the fixed gap $\gamma - \eta$.

Applying Lemma 1 gives the distribution-testing bound

$$q_{1/10}(\eta, \gamma; M) = O_{\gamma-\eta}(M/\log M) \quad (5)$$

for fixed $\eta < \gamma$ and all sufficiently large M .

4.3 From Testing to Compression

Theorem 2 (Compress by Test). *Fix $\eta \in [0, 1)$ and $\delta \in [0, 1/2)$ with $\gamma := 1 - 2\delta > \eta$, and fix $0 < \rho < 1/2$. Let $c \in \mathbb{N}$, and let q be a distribution-testing sample bound for these parameters on an alphabet of size 2^c ; that is, $q \geq q_\rho(\eta, 1 - 2\delta; 2^c)$ samples from each distribution suffice. Then for a uniformly random $f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}$,*

$$\Pr_f[C_{\text{CDS}}^{\eta, \delta}(f) \leq c] \leq 2^{2Nqc + \lceil \log(c+1) \rceil + 2 - (1 - H_2(\rho))N^2}. \quad (6)$$

In particular, if $2Nqc \leq \frac{1}{2}(1 - H_2(\rho))N^2$ and n is large, the right-hand side is $2^{-\Omega(N^2)}$.

Proof. Fix one deterministic test achieving the stated sample bound. We will describe each function with a cheap protocol by a local-message table and the test's error pattern. Let S be the set of f with $C_{\text{CDS}}^{\eta, \delta}(f) \leq c$, and fix $f \in S$ together with a protocol achieving it, padded so that $a + b = c$.

Sample q tapes under each secret and let T record all the local messages they induce, as in Theorem 1; this is $2Nqc$ bits. For each input (x, y) the table determines the q transcripts $(A(0, x, r_i^0), B(0, y, r_i^0))_{i \in [q]}$ and the q transcripts $(A(1, x, r_i^1), B(1, y, r_i^1))_{i \in [q]}$, which are independent samples from $D_{x,y}^0$ and from $D_{x,y}^1$ respectively, on an alphabet of size at most 2^c . Define

$$g_T(x, y) := \text{Test}\left(\left(A(0, x, r_i^0), B(0, y, r_i^0)\right)_i, \left(A(1, x, r_i^1), B(1, y, r_i^1)\right)_i\right).$$

This is computable from T , the split $a + b = c$, and (x, y) . By η -privacy and δ -correctness, the pair $(D_{x,y}^0, D_{x,y}^1)$ satisfies the promise of Definition 2 at every (x, y) , with the close case at 0-inputs and the far case at 1-inputs. Hence $\Pr_T[g_T(x, y) \neq f(x, y)] \leq \rho$ for every (x, y) .

The errors at different inputs are highly dependent — all inputs share the same tapes — but we only need their expected number. By linearity, $\mathbb{E}_T[d_H(g_T, f)] \leq \rho N^2$, so we may fix tapes with $d_H(g_T, f) \leq \rho N^2$.

Now encode f as the concatenation of (i) the split a , in $\lceil \log(c+1) \rceil$ bits; (ii) the table T , in $2Nqc$ bits; and (iii) the index of the error pattern $e := f \oplus g_T$ among all strings of Hamming weight at most ρN^2 , in $\lceil \log \sum_{j \leq \rho N^2} \binom{N^2}{j} \rceil \leq H_2(\rho)N^2 + 1$ bits. The decoder reads a and T , recomputes g_T at every input, decodes e and outputs $g_T \oplus e = f$. The encoding is injective on S and has length at most $\ell := 2Nqc + \lceil \log(c+1) \rceil + H_2(\rho)N^2 + 1$. There are fewer than $2^{\ell+1}$ binary strings of length at most ℓ , so $|S| < 2^{\ell+1}$. A uniformly random function has one of 2^{N^2} equally likely truth tables. Therefore,

$$\Pr_f[f \in S] \leq 2^{\ell+1-N^2} = 2^{2Nqc + \lceil \log(c+1) \rceil + 2 - (1-H_2(\rho))N^2},$$

which proves (6). $\square$

Corollary 1 (The exponent of the test is the exponent of the bound). *Fix η, δ as in Theorem 2, put $\gamma = 1 - 2\delta$ and $\rho = 1/10$, and write $q(M) := q_\rho(\eta, \gamma; M)$ for the corresponding distribution-testing sample complexity.*

1. *If $q_\rho(\eta, \gamma; M) \leq KM^\alpha$ for some $\alpha \in (0, 1]$, some $K \geq 1$, and all large M , then for $c_0 := \lfloor \frac{1}{\alpha}(n - \log n - L) \rfloor$ with $L := \log(2K/(\varepsilon\alpha))$ and $\varepsilon := \frac{1}{2}(1 - H_2(1/10)) > \frac{1}{4}$,*

$$\Pr_f[C_{\text{CDS}}^{\eta, \delta}(f) \leq c_0] \leq 2^{-\Omega(2^{2n})}.$$

2. *If instead $q_\rho(\eta, \gamma; M) \leq KM/\log M$, the same conclusion holds with $c_0 := \lfloor n - \log(2K/\varepsilon) \rfloor$.*

Proof. For (i), the choice of c_0 gives $c_0 \leq n/\alpha$, so $2Nq(2^{c_0})c_0 \leq 2KN2^{\alpha c_0}n/\alpha$, and this is at most εN^2 exactly when $\alpha c_0 \leq n - \log n - \log(2K/(\varepsilon\alpha))$, which is how c_0 was chosen. Apply Theorem 2. For (ii), the factor $\log M = c$ cancels the factor c in the table size: $2Nq(2^c)c \leq 2KN2^c$, which is at most εN^2 precisely when $c \leq n - \log(2K/\varepsilon)$. $\square$

4.4 Instantiations

Corollary 2 (Imperfect privacy, perfect correctness). *Let $0 \leq \eta < 1$ be fixed and $\alpha_\eta := 1 - \eta$. Then $q_{1/10}(\eta, 1; M) = O(\sqrt{M}/\alpha_\eta)$, and a random f satisfies*

$$C_{\text{CDS}}^{\eta, 0}(f) \geq 2n - 2\log n - 2\log \frac{1}{1 - \eta} - O(1)$$

with probability $1 - 2^{-\Omega(2^{2n})}$.

Here we can effectively reduce to the collision test of Theorem 1.

Proof. We may assume $\eta > 0$, the case $\eta = 0$ being Theorem 1.

The far case is unchanged: disjoint supports admit no collision. For the close case, suppose $\Delta(P, Q) \leq \eta$. The common-region decomposition (see Vadhan's thesis for elementary presentation [Vad99, proof of Lemma 3.1.8, p. 37]) gives the existence of a distribution D shared by P and Q with weight $1 - \Delta(P, Q) \geq \alpha_\eta$. In other words, we can write P and Q as mixtures involving D :

$$P = \alpha_\eta D + (1 - \alpha_\eta)P', \quad Q = \alpha_\eta D + (1 - \alpha_\eta)Q'$$

for some distributions P', Q' .

Let $u := \lceil \sqrt{M} \rceil$ and draw $q_0 := \lceil 4u/\alpha_\eta \rceil$ samples from each side. The number of samples coming from the D -component is binomially distributed with mean at least $4u$ in each batch, and the two counts are independent because the two batches toss disjoint coins; a Chernoff bound gives $\Pr[\text{count} < u] \leq e^{-9u/8}$. Even if we condition on both counts being at least u , the first u D -samples on each side are still independent draws from D , whose support has size at most $M \leq u^2$. Proposition 1 therefore produces a collision among them — and hence among all the samples — with probability at least $1/4$, so

$$\Pr[\text{collision}] \geq \frac{1}{4}(1 - e^{-9/8})^2 > \frac{1}{10}.$$

Lemma 2 with 22 blocks gives error $(9/10)^{22} < 1/10$ and $q = O(\sqrt{M}/\alpha_\eta)$. Apply Corollary 1(i) with $\alpha = 1/2$ and $K = O(1/\alpha_\eta)$: there $L = \log \frac{1}{1-\eta} + O(1)$, and the factor $1/\alpha = 2$ doubles it in the final bound. $\square$

Corollary 3 (Perfect privacy, imperfect correctness). *Let $0 \leq \delta < 1/2$ be fixed. Then a random f satisfies*

$$C_{\text{CDS}}^{0,\delta}(f) \geq \frac{3}{2}(n - \log n) - O_\delta(1)$$

with probability $1 - 2^{-\Omega(2^{2n})}$.

Proof. With $\gamma = 1 - 2\delta > 0$ fixed, (4) gives the distribution-testing bound $q_{1/10}(0, \gamma; M) = O_\delta(M^{2/3})$. Apply Corollary 1(i) with $\alpha = 2/3$. $\square$

Corollary 4 (Both imperfect). *Let $\eta \in [0, 1)$ and $\delta \in [0, 1/2)$ be fixed with*

$$\eta < 1 - 2\delta. \tag{7}$$

Then a random f satisfies $C_{\text{CDS}}^{\eta,\delta}(f) \geq n - O_{\eta,\delta}(1)$ with probability $1 - 2^{-\Omega(2^{2n})}$.

Proof. Put $\gamma := 1 - 2\delta$. Since $\gamma - \eta > 0$ is fixed, the tolerant-testing bound in (5) gives $q_{1/10}(\eta, \gamma; M) = O_{\eta,\delta}(M/\log M)$. Apply Corollary 1(ii). $\square$

Remark 1 (Elementary version of Corollary 4). We note that invoking Valiant and Valiant's test is unnecessary, and there is an elementary proof that recovers a similar result. Set $\tau := 1 - 2\delta - \eta > 0$. Let $\hat{P}$ be the empirical distribution of q samples from P . Cauchy–Schwarz gives $\mathbb{E}[\Delta(\hat{P}, P)] \leq \frac{1}{2}\sqrt{M/q}$, and triangle inequality

$$|\Delta(\hat{P}, \hat{Q}) - \Delta(P, Q)| \leq \Delta(\hat{P}, P) + \Delta(\hat{Q}, Q).$$

Taking $q = \lceil 900M/\tau^2 \rceil$ and applying Markov's inequality yields an additive- $\tau/3$ estimate with probability at least $9/10$, hence a test with $q = O_{\eta,\delta}(M)$ samples. Corollary 1(i) with $\alpha = 1$ then gives $C_{\text{CDS}}^{\eta,\delta}(f) \geq n - \log n - O_{\eta,\delta}(1)$.

Acknowledgments

The author is supported in part by the U.S. National Science Foundation through CAREER Award No. 2443735 and Award No. 2552209, the latter under the joint NSF–United States–Israel Binational Science Foundation (BSF) program. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

AI disclosure. The author used OpenAI’s Codex to assist with mathematical arguments and proof drafts, literature searches, exposition, technical review, and writing this disclosure statement. :) The author is responsible for the results, proofs, citations, and final presentation.

References

- [AARV21] Benny Applebaum, Barak Arkis, Pavel Raykov, and Prashant Nalini Vasudevan. Conditional disclosure of secrets: Amplification, closure, amortization, lower-bounds, and separations. *SIAM Journal on Computing*, 50(1):32–67, 2021. <https://doi.org/10.1137/18M1217097>.
- [AHMS20] Benny Applebaum, Thomas Holenstein, Manoj Mishra, and Ofer Shayeitz. The communication complexity of private simultaneous messages, revisited. *Journal of Cryptology*, 33(3):917–953, 2020. Conference version: EUROCRYPT 2018. <https://doi.org/10.1007/s00145-019-09334-y>. Full version: <https://eprint.iacr.org/2018/144>.
- [AV21] Benny Applebaum and Prashant Nalini Vasudevan. Placing conditional disclosure of secrets in the communication complexity universe. *Journal of Cryptology*, 34(2):11, 2021. <https://doi.org/10.1007/s00145-021-09376-1>. Conference version: ITCS 2019, Article 4, <https://doi.org/10.4230/LIPIcs.ITCS.2019.4>.
- [Can20] Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. <https://doi.org/10.4086/toc.gs.2020.009>.
- [Can22] Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends in Communications and Information Theory*, 19(6):1032–1198, 2022. <https://doi.org/10.1561/0100000114>.
- [CDVV14] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203, 2014. <https://doi.org/10.1137/1.9781611973402.88>. Author copy: <https://www.cs.purdue.edu/homes/pvaliant/closeness.pdf>.
- [FKN94] Uri Feige, Joe Kilian, and Moni Naor. A minimal model for secure computation. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, pages 554–563, 1994. <https://doi.org/10.1145/195058.195408>.
- [GIKM00] Yael Gertner, Yuval Ishai, Eyal Kushilevitz, and Tal Malkin. Protecting data privacy in private information retrieval schemes. *Journal of Computer and System Sciences*, 60(3):592–629, 2000. <https://doi.org/10.1006/jcss.1999.1689>.
- [GKW15] Romain Gay, Iordanis Kerenidis, and Hoeteck Wee. Communication complexity of conditional disclosure of secrets and attribute-based encryption. In *Advances in Cryptology — CRYPTO 2015, Part II*, pages 485–502, 2015. https://doi.org/10.1007/978-3-662-48000-7_24.
- [LVW17] Tianren Liu, Vinod Vaikuntanathan, and Hoeteck Wee. Conditional disclosure of secrets via non-linear reconstruction. In *Advances in Cryptology — CRYPTO 2017, Part I*, pages 758–790, 2017. https://doi.org/10.1007/978-3-319-63688-7_25.
- [Vad99] Salil Pravin Vadhan. *A Study of Statistical Zero-Knowledge Proofs*. PhD thesis, Massachusetts Institute of Technology, 1999. Minor revision, August 2000. <https://people.seas.harvard.edu/~salil/research/phdthesis.pdf>.
- [VV11] Gregory Valiant and Paul Valiant. The power of linear estimators. In *52nd Annual IEEE Symposium on Foundations of Computer Science*, pages 403–412, 2011. <https://doi.org/10.1109/FOCS.2011.81>. Full version (including Theorem 4, Appendix C): https://theory.stanford.edu/~valiant/papers/VV_LinEst.pdf.

A Proof of the Birthday Bound

Proof of Proposition 1. Let $p_\omega := D(\omega)$. Define the following generalization of the collision probability of D for $k \in \mathbb{N}$:

$$s_k := \sum_{\omega \in \Omega} p_\omega^k.$$

Let $I_{i,j}$ be an indicator for a collision $X_i = Y_j$,

$$I_{i,j} := \mathbf{1}[X_i = Y_j].$$

Let W count the number of collisions between X samples and Y samples:

$$W := \sum_{i,j \in [t]} I_{i,j}, \quad \mu := \mathbb{E}[W] = t^2 s_2.$$

It suffices to show $\Pr[W > 0] \geq 1/4$.

By Cauchy–Schwarz,

$$1 = \left(\sum_{\omega} p_\omega \right)^2 \leq |\Omega| \sum_{\omega} p_\omega^2,$$

so

$$s_2 \geq \frac{1}{|\Omega|} \geq \frac{1}{t^2} \quad \implies \quad \mu = t^2 s_2 \geq 1.$$

Additionally,

$$\begin{aligned} s_3 &= \sum_{\omega} p_\omega^3 \leq \left(\max_{\omega} p_\omega \right) \sum_{\omega} p_\omega^2 \\ &= \sqrt{\max_{\omega} p_\omega^2} \sum_{\omega} p_\omega^2 \\ &\leq \sqrt{\sum_{\omega} p_\omega^2} \sum_{\omega} p_\omega^2 = s_2^{3/2}. \end{aligned}$$

Hence $t^3 s_3 \leq \mu^{3/2}$.

Expanding the raw second moment, with every index below ranging over $[t]$,

$$\begin{aligned} \mathbb{E}[W^2] &= \mathbb{E} \left[\sum_{i,j,k,\ell} I_{i,j} I_{k,\ell} \right] = \sum_{i,j,k,\ell} \mathbb{E}[I_{i,j} I_{k,\ell}] \\ &= \sum_{\substack{i=k \\ j=\ell}} \mathbb{E}[I_{i,j} I_{k,\ell}] + \sum_{\substack{i=k \\ j \neq \ell}} \mathbb{E}[I_{i,j} I_{k,\ell}] \\ &\quad + \sum_{\substack{i \neq k \\ j=\ell}} \mathbb{E}[I_{i,j} I_{k,\ell}] + \sum_{\substack{i \neq k \\ j \neq \ell}} \mathbb{E}[I_{i,j} I_{k,\ell}] \end{aligned}$$

$$\begin{aligned}
&= \sum_{i,j} \mathbb{E}[I_{i,j}^2] + \sum_{\substack{i,j,\ell \\ j \neq \ell}} \mathbb{E}[\mathbf{1}[X_i = Y_j = Y_\ell]] \\
&\quad + \sum_{\substack{i,k,j \\ i \neq k}} \mathbb{E}[\mathbf{1}[X_i = X_k = Y_j]] + \sum_{\substack{i \neq k \\ j \neq \ell}} \mathbb{E}[I_{i,j}] \mathbb{E}[I_{k,\ell}] \\
&= \sum_{i,j} \mathbb{E}[I_{i,j}] + 2 \sum_{\substack{i,j,\ell \\ j \neq \ell}} \mathbb{E}[\mathbf{1}[X_i = Y_j = Y_\ell]] \\
&\quad + \sum_{\substack{i \neq k \\ j \neq \ell}} \mathbb{E}[I_{i,j}] \mathbb{E}[I_{k,\ell}].
\end{aligned}$$

Here the first sum has t^2 terms, each with expectation s_2 . Each of the two sums sharing exactly one index has $t^2(t-1)$ terms, each with expectation s_3 . The final sum has $t^2(t-1)^2$ terms. In that sum the two indicators depend on four distinct sample variables, so they are independent and their product has expectation s_2^2 . Consequently,

$$\begin{aligned}
\mathbb{E}[W^2] &= t^2 s_2 + 2t^2(t-1)s_3 + t^2(t-1)^2 s_2^2 \\
&\leq t^2 s_2 + 2t^3 s_3 + t^4 s_2^2 \\
&\leq \mu + 2\mu^{3/2} + \mu^2 \\
&= (\mu + \sqrt{\mu})^2 \\
&\leq 4\mu^2.
\end{aligned}$$

The last line follows from our observation that $\mu \geq 1$.

On the other hand, by Cauchy–Schwarz,

$$\mathbb{E}[W]^2 = (\mathbb{E}[W \mathbf{1}[W > 0]])^2 \leq \mathbb{E}[W^2] \Pr[W > 0].$$

So,

$$\Pr[W > 0] \geq \frac{\mathbb{E}[W]^2}{\mathbb{E}[W^2]} \geq \frac{\mu^2}{4\mu^2} = \frac{1}{4}.$$

□