Interactive Fingerprinting Codes and the Hardness of Preventing False Discovery

Thomas Steinke* Jonathan Ullman†

October 5, 2014

Abstract

We show a tight bound on the number of adaptively chosen statistical queries that a computationally efficient algorithm can answer accurately given n samples from an unknown distribution. A statistical query asks for the expectation of a predicate over the underlying distribution, and an answer to a statistical query is accurate if it is "close" to the correct expectation over the distribution. This question was recently considered by Dwork et al. [DFH+14], who showed that $\tilde{\Omega}(n^2)$ queries can be answer efficiently, and also by Hardt and Ullman [HU14], who showed that answering $\tilde{O}(n^3)$ queries is computationally hard. We close the gap between the two bounds by proving a new, nearly-optimal hardness result. Specifically, we show that, under a standard hardness assumption, there is no computationally efficient algorithm that given n samples from an unknown distribution can give valid answers to $O(n^2)$ adaptively chosen statistical queries. An implication of our results is that computationally efficient algorithms for answering arbitrary, adaptively chosen statistical queries may as well be *differentially private*. We obtain our results via an optimal construction of a new combinatorial object that we call an *interactive fingerprinting code*, which may be of independent interest.

*Harvard University School of Engineering and Applied Sciences. Supported by NSF grant CCF-1116616. Email: tsteinke@seas.harvard.edu.

†Harvard University Center for Research on Computation and Society and Columbia University. Supported by NSF Grant CNS-1237235 and a Simons Society of Fellows Junior Fellowship. Email: jullman@cs.columbia.edu.

Contents

1 Introduction

Empirical research is commonly done by asking multiple "queries" (e.g., summary statistics, hypothesis tests, or learning algorithms) on a finite sample drawn from some population. "False discovery" occurs if the outcome of the queries on the sample does not reflect the population. For example, if a certain irrelevant gene is believed to be predictive for cancer based on the sample. For decades statisticians have been devising methods for preventing false discovery, such as the "Bonferroni correction" [Bon36, Dun61] and the widely used and highly influential method of Benjamini and Hochberg [BH95] for controlling the "false discovery rate."

Nevertheless, false discovery persists across all empirical sciences, and both popular and scientific articles report on an increasing number of invalid research findings. Typically false discovery is attributed to misuse of statistics. However, another possible explanation—recently proposed by Dwork et al. [DFH+14] and Hardt and Ullman [HU14]—is that methods for preventing false discovery do not address the fact that modern data analysis is inherently *adaptive*. The queries made, the experimental setup, and the selection and tuning of the algorithms all depend on previous interactions with the data. [DFH+14] gave methods for preventing false discovery in some cases, while [HU14] showed contrasting hardness results suggesting that there may be an inherent computational barrier to preventing false discovery. In this work we give a new, nearly optimal hardness result, showing that the algorithms of Dwork et al. [DFH+14] are nearly optimal in the worst case.

These results are formalized in Kearns' *statistical-query (SQ) model* [Kea93]. In the SQ model, there is an algorithm called the *oracle* that gets access to n samples from an unknown distribution $\mathcal{D}$ over some finite universe $\{0,1\}^d$, where the parameter d is the dimensionality of the distribution. The oracle must answer *statistical queries* about $\mathcal{D}$. A statistical query q is specified by a predicate $p\colon \mathcal{X} \to \{0,1\}$ and the answer to a statistical query is $q(\mathcal{D}) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[p(x)]$.

The oracle's answer a to a query q is *accurate* if $|a - q(\mathcal{D})| \leq \alpha$ with high probability. Importantly, the goal of the oracle is to provide answers that "generalize" to the underlying distribution rather than answers that are specific to the sample. The latter is easy to achieve by outputting the empirical average of the query predicate on the sample.

The analyst makes a sequence of queries $q^1, q^2, \ldots, q^k$ to the oracle, which responds with answers $a^1, a^2, \ldots, a^k$. In the adaptive setting, the analyst receives each answer a^i to q^i before the next query is asked, so q^{i+1} may depend on $q^1, a^1, q^2, a^2, \ldots, q^i, a^i$. We say the oracle is *accurate* given n samples for k adaptively chosen queries, if when given n samples from an arbitrary distribution $\mathcal{D}$, the oracle accurately responds to any adaptive analyst that makes at most k queries. A computationally efficient oracle answers each query in time polynomial in n and d.[1]

When the queries $q^1, q^2, \ldots, q^k$ are specified *non adaptively* (i.e. independent of $a^1, a^2, \ldots, a^k$), then the empirical average of each query on the sample is accurate with high probability as long as $k \leq 2^{o(n)}$. However, this guarantee no longer holds in the interactive setting and a more sophisticated oracle is required.

The results of Dwork et al. [DFH+14] gave a surprising way to construct an accurate oracle, by showing that any oracle that satisfies *differential privacy* [DMNS06] and provides accurate answers with respect to the sample, also provides accurate answers with respect to the underlying distribution. Using known constructions of differentially private algorithms, they obtain a computationally efficient algorithm that is accurate for $\tilde{\Omega}(n^2)$ queries, and a computationally inefficient algorithm that is accurate for nearly 2^n queries.

[1] We assume that the analyst asks queries that can be evaluated on the sample in polynomial time.

It was already known that computationally efficient differentially private algorithms could not answer more than $\tilde{O}(n^2)$ queries [Ull13], but differentially privacy may not be necessary at all. Unfortunately, [HU14] showed that *no* computationally efficient algorithm could answer too many more queries. Specifically, assuming the existence of one-way functions, there is no computationally efficient oracle that answers more than $\tilde{O}(n^3)$ queries.

In this work we resolve this gap almost completely, and show that, assuming the existence of one-way functions, there is no efficient oracle that answers $O(n^2)$ queries. Conceptually, our result gives further evidence that there may be an inherent computational barrier to preventing false discovery in interactive data analysis. It also shows that when answering arbitrary adaptively chosen statistical queries, one cannot do better than to use a differentially private algorithm in the worst case. We believe it is an intriguing open question to see whether this sort of equivalence holds in more restricted settings. Finally, as a consequence of our results, we obtain a dichotomy for privacy preserving data analysis in the adaptive setting. There is a computationally efficient differentially private algorithm that answers $\tilde{\Omega}(n^2)$ arbitrary adaptive queries, but, assuming the existence of one-way functions, every computationally efficient algorithm that answers $O(n^2)$ arbitrary adaptive queries is *blatantly non private*.

To prove our result, we modify the adversary used in [HU14]. First, we abstract the combinatorial properties required by their adversary into a new combinatorial object that we call an *interactive fingerprinting code*. Then, we give a nearly optimal construction of interactive fingerprinting codes, which is the main technical contribution of our work.

1.1 Discussion of Results

Our main result is the following nearly optimal hardness result for preventing false discovery in interactive data analysis.

Theorem 1.1 (Informal)**.** *Assuming the existence of one-way functions, there is no computationally efficient oracle that given n samples is accurate on $O(n^2)$ adaptively chosen queries.*

As in [HU14], our hardness result applies whenever the dimensionality of the data grows with the sample size faster than logarithmically so that 2^d is no longer polynomial in n.[2] This requirement is rather mild, and is also necessary. If $n \gg 2^d$ then the empirical distribution of the n samples will be close to the underlying distribution in statistical distance, so every statistical query can be answered accurately given the sample. Thus, the dimensionality of the data has a major effect on the hardness of the problem. [HU14] also showed that if the dimensionality is much larger than n, then we cannot even hope for a *computationally unbounded* oracle that provides accuracy on adaptive queries. We obtain a nearly optimal version of that result.

Theorem 1.2 (Informal)**.** *There is no oracle (even a computationally unbounded one) that given n samples in dimension $d = O(n^2)$ is accurate on $O(n^2)$ adaptively chosen queries.*

This result should be contrasted with the aforementioned result of Dwork et al. [DFH+14] showing that if $d \ll n^2$ then there is an oracle that runs in time polynomial in n and 2^d and accurately answers nearly 2^n adaptive queries.

[2] This is under the stronger, but still standard, assumption that exponentially hard one-way functions exist.

1.1.1 Techniques

The structure of our proof is rather simple, and closely follows the framework in [HU14]. We will design a challenge distribution $\mathcal{D}$ and a computationally efficient adaptive analyst $\mathcal{A}$ who knows $\mathcal{D}$. If any computationally efficient oracle $\mathcal{O}$ is given n samples $S = \{x_1, \ldots, x_n\}$ drawn from $\mathcal{D}$, then our analyst $\mathcal{A}$ can use the answers of $\mathcal{O}$ to reconstruct the set S. Using this information, the adversary can construct a query on which S is not representative of $\mathcal{D}$.

Our adversary $\mathcal{A}$ and the distribution $\mathcal{D}$, like that of [HU14], is built from a combinatorial object with a computational "wrapper." The computational wrapper is cryptographic queries which "hide" information from the oracle $\mathcal{O}$. The combinatorial object is what we call an *interactive fingerprinting code*, which we introduce.

Interactive fingerprinting codes (IFPCs) are an abstraction of the technique in [HU14]. They generalize *fingerprinting codes*, which were introduced by Boneh and Shaw [BS98] for the problem of watermarking digital content. Our main contribution is to define and construct IFPCs.

An interactive fingerprinting code $\mathcal{F}$ is an efficient interactive algorithm that defeats any adversary $\mathcal{P}$ (called the *pirate*) with high probability in the following game. The adversary $\mathcal{P}$ picks $S \subset [N]$ unknown to $\mathcal{F}$. The goal of $\mathcal{F}$ is to identify S by making ℓ interactive queries to $\mathcal{P}$. $\mathcal{F}$ specifies each query by a vector $c \in \{\pm 1\}^N$. In response, the adversary $\mathcal{P}$ must simply output $a \in \{\pm 1\}$ such that $a = c_i$ for some $i \in [N]$. However, the adversary is restricted to only see c_i for $i \in S$. At any time, $\mathcal{F}$ may *accuse* some $i \in [N]$. If $i \in S$ is accused, then i is removed from S (i.e. $S \leftarrow S \backslash \{i\}$), thereby further restricting $\mathcal{P}$. If $i \notin S$ is accused, then this is referred to as a *false accusation*. To win, the interactive fingerprinting code $\mathcal{F}$ must identify all of S, without making "too many" false accusations.

The difference between interactive and non interactive fingerprinting codes is that a non interactive fingerprinting code must give all ℓ queries to $\mathcal{P}$ at once, but is (necessarily) only required to identify one $i \in S$.

[HU14] implicitly construct an interactive fingerprinting code with $\ell = \tilde{O}(N^3)$ by concatenating N independent copies of a non interactive fingerprinting code. We give a construction of an interactive fingerprinting codes using the optimal $O(N^2)$ queries based on the construction of non interactive fingerprinting codes by [Tar08].

Theorem 1.3 (Informal)**.** *For every N, there exists an interactive fingerprinting code with $\ell = O(N^2)$ that, except with negligible probability, makes at most $N/1000$ false accusations.*

This result suffices for our applications, but our construction is somewhat more general and has several additional parameters, which we detail in Section 2.

1.2 Applications to Data Privacy

The adversary used to show hardness of preventing false discovery is effectively carrying out a *reconstruction attack* against the database of samples. Roughly, if there is an adversary who can reconstruct the set of samples S from the oracle's answers, then the oracle is said to be "blatantly non-private"—it reveals essentially all of the data it holds, and so cannot guarantee any reasonable notion of privacy to the owners of the data. Since the seminal work of Dinur and Nissim [DN03], such reconstruction attacks have been used to establish strong limitations on the accuracy of privacy-preserving oracles.

Using interactive fingerprinting codes, combined with the framework of [HU14], we obtain the following results. In both cases, [HU14] show similar results, in which our $O(n^2)$ bounds are replaced with $\tilde{O}(n^3)$.

Theorem 1.4 (Informal)**.** *Assuming the existence of one-way functions, every computationally efficient oracle that, given n samples, is accurate on $O(n^2)$ adaptively chosen queries is blatantly non private.*

Theorem 1.4 should be compared with the result in [Ull13], which showed that any computationally efficient oracle that, given n samples, is accurate for $\tilde{O}(n^2)$ *non-adaptively chosen* queries cannot satisfy the strong guarantee of "differential privacy" [DN03, DMNS06]. Theorem 1.4 shows that, in the adaptive setting, we can obtain a stronger privacy violation using fewer queries than [Ull13].

Theorem 1.5 (Informal)**.** *Every (possibly computationally unbounded) oracle that, given n samples in dimension $d = O(n^2)$, is accurate on $O(n^2)$ adaptively chosen queries is blatantly non private.*

Theorem 1.5 should be compared with the result in [BUV14] that showed any (possibly computationally unbounded) oracle that answers a *fixed* family of $\tilde{O}(n^2)$ simple queries in dimension $d = \tilde{O}(n^2)$ cannot satisfy differential privacy.

In contrast with Theorems 1.4 and 1.5, the well-known result of [DMNS06] shows that there is an efficient differentially private algorithm that answers $\tilde{\Omega}(n^2)$ adaptively chosen queries. Our results show that, in the adaptive setting, there is a sharp threshold for the number of queries where, below this threshold, the strong notation of differential privacy can be achieved and, above this threshold, even minimal notions of privacy are unachievable.

1.3 Related Work

Our work and [HU14] build on techniques for attacking allegedly privacy preserving algorithms. These works showed a surprising tension between methods for *secure content-distribution* and privacy-preserving algorithms. This connection first appeared in the work of Dwork, Naor, Reingold, Rothblum, and Vadhan [DNR+09], who showed that the existence of *traitor-tracing schemes* [CFN94] implies hardness results for differential privacy. This connection was expanded and strengthened in [Ull13], which introduced the use of fingerprinting codes in the context of differential privacy, and used them to prove nearly optimal hardness results for certain problems in differential privacy. Bun et al. [BUV14] showed that fingerprinting codes can be used to prove nearly-optimal information-theoretic lower bounds for differential privacy, which established fingerprinting codes as the key information-theoretic object underlying lower bounds in differential privacy.

Interactive fingerprinting codes are similar in spirit to the work of Fiat and Tassa [FT01] on *dynamic traitor-tracing*. They too used the power of adaptivity to reconstruct the entire set of samples, as opposed to only one sample. Technically their results are incomparable to ours, and their techniques do not suffice in our setting. Specifically, they allowed codewords over a large alphabet, $c \in [N]^N$, and achieve a length of $\tilde{O}(N)$. This large alphabet generalization is fundamentally different, and does not lead to hardness results for answering adaptive queries. This is inherent, since, by [DFH+14], there is no hardness result for answering $\tilde{O}(N)$ queries.[3]

The algorithms of Dwork et al. [DFH+14] rely on known differentially private mechanisms for answering adaptive statistical queries. Recently, [Ull14] showed how to design differentially private mechanisms for answering exponentially many adaptively chosen queries

[3]This problem is not merely an artifact of restricting the data to distributions over $\{\pm 1\}$. Even if we were to consider more general distributions over $[N]$, the methods of Fiat and Tassa [FT01] inherently do not suffice. Queries over $[N]$ are equivalent to queries over $\{\pm 1\}^{\log N}$, whereas the large alphabet is fundamental to the analysis in Fiat and Tassa and cannot be straightforwardly replaced by the binary alphabet.

from the richer class of *convex empirical risk minimization queries*. By the results of Dwork et al. [DFH+14], this algorithm is also a (computationally inefficient) oracle that is accurate for exponentially many adaptively chosen convex empirical risk minimization queries.

Organization In Section 2 we define and construct interactive fingerprinting codes, the main new technical ingredient we use to establish our results. In Sections 3 and 4 we show how interactive fingerprinting codes can be used to obtain hardness results for preventing false discovery and blatant non privacy, respectively. The definition of interactive fingerprinting codes is contained in Section 2.1 and is necessary for Sections 3 and 4, but the remainder of Section 2 and Sections 3 and 4 can be read in either order.

2 Interactive Fingerprinting Codes

In order to motivate the definition of interactive fingerprinting codes, it will be helpful to review the motivation for standard, non-interactive fingerprinting codes.

Fingerprinting codes were introduced by Boneh and Shaw [BS98] for the problem of watermarking digital content (such as a movie or a piece of software). Consider a company that distributes some content to N users. Some of the users may illegally distribute copies of the content. To combat this, the company gives each user a unique version of the content by adding distinctive "watermarks" to it. Thus, if the company finds an illegal copy, it can be traced back to the user who originally purchased it. Unfortunately, users may be able to remove the watermarks. In particular, a coalition of users may combine their copies in a way that mixes or obfuscates the watermarks. A fingerprinting code ensures that, even if up to n users collude to combine their codewords, an illegal copy can be still be traced to at least one of the users.

Formally, every user $i \in [N]$ is given a codeword $(c_i^1, c_i^2, \ldots, c_i^\ell) \in \{\pm 1\}^\ell$ by the fingerprinting code, which represents the combination of watermarks in that user's copy. A subset $S \subset [N]$ of at most n users can arbitrarily combine their codewords to create a "pirate codeword" $a = (a^1, a^2, \ldots, a^\ell) \in \{\pm 1\}^\ell$. The only constraint is so-called *consistency*—for every $j \in [\ell]$, if, for every colluding user $i \in S$, we have $c_i^j = b$, then $a^j = b$. That is to say, if each of the colluding users receives the same watermark, then their combined codeword must also have that watermark. Given a, the fingerprinting code must be able to trace at least one user $i \in S$. Tardos [Tar08] constructed optimal fingerprinting codes with $\ell = O(n^2 \log N)$.

A key drawback of fingerprinting codes is that we can only guarantee that a single user $i \in S$ is traced. This is inherent, as setting the pirate codeword a to be the codeword of a single user prevents any other user from being identified. We will see that this can be circumvented by moving to an interactive setting.

Suppose the company is instead distributing a *stream* of content (such as a TV series) to N users—that is, the content is not distributed all at once and the illegal copies are obtained whilst the content is being distributed (e.g. the episodes of the TV series appear on the internet before the next episode is shown). Again, the content is watermarked so that each user receives a unique stream and a subset $S \subset [N]$ of at most n users combine their streams and distribute an illegal stream. The company obtains the illegal stream and uses this to trace the colluding users S. As soon as the company can identify a colluding user $i \in S$, that user's stream is terminated (e.g. their subscription is cancelled). This process continues until every $i \in S$ has been traced and the distribution of illegal copies ceases.

Another twist on fingerprinting codes is robustness [BUV14]. Suppose that the consistency constraint only holds for $(1-\beta)\ell$ choices of $j \in [\ell]$. That is to say, the colluding users can somehow remove a β fraction of the watermarks. [BUV14] showed how to modify the Tardos fingerprinting code to be robust to a small constant fraction of inconsistencies. In this work, we show that robustness to any $\beta < 1/2$ fraction of inconsistencies can be achieved.

2.1 Definition and Existence

We are now ready to formally define interactive fingerprinting codes. To do so we make use of the following game between an adversary $\mathcal{P}$ and the fingerprinting code $\mathcal{F}$. Both $\mathcal{P}$ and $\mathcal{F}$ may be stateful. For a given execution of $\mathcal{F}$, we let $C \in \{\pm 1\}^{N \times \ell}$ be the matrix with columns $c^1, \ldots, c^\ell$

$\mathcal{P}$ selects a subset of users $S^1 \subseteq [N]$ of size at most n, unknown to $\mathcal{F}$
For $j = 1, \ldots, \ell$:
  $\mathcal{F}$ outputs a column vector $c^j \in \{\pm 1\}^N$
  Let $c^j_{S^j} \in \{\pm 1\}^{|S^j|}$ be the restriction of c^j to coordinates in S^j, which is given to $\mathcal{P}$
  $\mathcal{P}$ outputs a^j, which is given to $\mathcal{F}$
  $\mathcal{F}$ accuses a (possibly empty) set of users $I^j \subseteq [N]$. Let $S^{j+1} = S^j \setminus I^j$.

Figure 1: $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}]$

and let $a \in \{\pm 1\}^\ell$ be the vector with entries $a^1, \ldots, a^\ell$. We want to construct the fingerprinting code so that, if a is consistent, then the tracer succeeds in recovering every user in S. For convenience, we will define the notation θ^j to be the number of rounds $1, \ldots, j$ in which a^j is not consistent with c^j. Formally, for a given execution of $\mathcal{F}$,

$$\theta^j = \left| \left\{ 1 \le k \le j \;\middle|\; \nexists\, i \in [N],\ a^k = c^k_i \right\} \right|.$$

Using this notation, a is β-consistent if $\theta^\ell \le \beta\ell$. We also define the notation ψ^j to be the number of users in $I^1, \ldots, I^j$ who are falsely accused (i.e. not in the coalition S^1). Formally,

$$\psi^j = \left| \left(\bigcup_{1 \le k \le j} I^k \right) \setminus S^1 \right|.$$

Using this notation, we require $\psi^\ell \le \delta N$ - that is, the tracing algorithm does not make too many false accusations.

Definition 2.1 (Interactive Fingerprinting Codes)**.** We say that an algorithm $\mathcal{F}$ is an *n-collusion-resilient interactive fingerprinting code of length ℓ for N users robust to a β fraction of errors with failure probability ε and false accusation probability δ* if for every adversary $\mathcal{P}$, it holds that

$$\mathop{\mathbb{P}}_{\mathsf{IFPC}_{N,n,\ell}[\mathcal{P},\mathcal{F}]} \left[\left(\theta^\ell \le \beta\ell \right) \vee \left(\psi^\ell > \delta N \right) \right] \le \varepsilon$$

The length ℓ may depend on $N, n, \beta, \varepsilon, \delta$.

The constraint $\psi^\ell \le \delta N$ is called *soundness*—the interactive fingerprinting code should not make (many) false accusations. The constraint $\theta^\ell > \beta\ell$ is called *completeness*—the interactive fingerprinting code should force the adversary $\mathcal{P}$ to be inconsistent. Although it may seem strange that we make no reference to recovering the coalition S^1, notice that if $S^j \neq \emptyset$, then $\mathcal{P}$ can easily be consistent. Therefore, if the pirate cannot be consistent, it must be the case that $S^j = \emptyset$ for some j, meaning all of S^1 has been accused.

In the remainder of this section, we give a construction of interactive fingerprinting codes, and establish the following theorem.

Theorem 2.2 (Existence of Interactive Fingerprinting Codes)**.** *For every $1 \le n \le N$, $0 \le \beta < 1/2$, and $0 < \delta \le 1$, there is a n-collusion-resilient interactive fingerprinting code of length ℓ for N users robust to a β fraction of errors with failure probability $\varepsilon \le \min\{\delta N, 2^{-\Omega(\delta N)}\} + \delta^{\Omega\left(\left(\frac{1}{2}-\beta\right)n\right)}$ and false accusation probability δ for*

$$\ell = O\left(\frac{n^2 \log(1/\delta)}{\left(\frac{1}{2}-\beta\right)^4}\right).$$

The expression for the failure probability ε is a bit mysterious. To interpret it, we fix $\beta = 1/2 - \Omega(1)$ and consider two parameter regimes: $\delta N \ll 1$ and $\delta N \gg 1$. In the traditional parameter regime for fingerprinting codes $\delta N = \varepsilon' \ll 1$, and so no users are falsely accused. Then our fingerprinting code has length $O(n^2 \log(N/\varepsilon'))$ and a failure probability of ε'. However, if we are willing to tolerate falsely accusing a small constant fraction of users, then we can set, for example, $\delta N = .01N$, and our fingerprinting code will have length $O(n^2)$ and failure probability $2^{-\Omega(n)}$.

2.2 The Construction

Our construction and analysis is based on the optimal (non interactive) fingerprinting codes of Tardos [Tar08], and the robust variant by Bun et al. [BUV14]. The code is essentially the same, but columns are generated and shown to the adversary one at a time, and tracing is modified to identify users interactively.

We begin with some definitions and notation. For $0 \le a < b \le 1$, let $D_{a,b}$ be the distribution with support (a,b) and probability density function $\mu(p) = C_{a,b}/\sqrt{p(1-p)}$, where $C_{a,b}$ is a normalising constant.[4] For $\alpha, \zeta \in (0, 1/2)$, let $\overline{D}_{\alpha,\zeta}$ be the distribution on $[0,1]$ that returns a sample from $D_{\alpha,1-\alpha}$ with probability $1-2\zeta$ and 0 or 1 each with probability ζ.

For $p \in [0,1]$, let $c \sim p$ denote that $c \in \{\pm 1\}$ is drawn from the distribution with $\mathbb{P}[c = 1] = p$ and $\mathbb{P}[c = -1] = 1 - p$. Let $c_{1\cdots n} \sim p$ denote that $c \in \{\pm 1\}^n$ is drawn from a product distribution in which $c_i \sim p$ independently for all $i \in [n]$.

Define $\phi^p : \{\pm 1\} \to \mathbb{R}$ by $\phi^0(c) = \phi^1(c) = 0$ and, for $p \in (0,1)$, $\phi^p(1) = \sqrt{(1-p)/p}$ and $\phi^p(-1) = -\sqrt{p/(1-p)}$. The function ϕ^p is chosen so that $\phi^p(c)$ has mean 0 and variance 1 when $c \sim p$.

The fingerprinting code $\mathcal{F}$ is defined in Figure 2. In addition to the precise setting of parameters, we have given asymptotic bounds to help follow the analysis. We now analyze $\mathcal{F}$ and establish Theorem 2.2. The proof of Theorem 2.2 is split into Theorems 2.7 and 2.17. For convenience, define $I = \bigcup_{j \in [\ell]} I^j$.

[4] To sample from $D_{a,b}$, first sample $\varphi \in (\sin^{-1}(\sqrt{a}), \sin^{-1}(\sqrt{b}))$ uniformly, then output $\sin^2(\varphi)$ as the sample.

Given parameters $1 \le n \le N$ and $0 < \delta, \beta < 1/2$
Set parameters:

$$\begin{aligned}
\alpha &= \frac{\left(\frac{1}{2}-\beta\right)}{4n} && \ge \Omega\left(\frac{\left(\frac{1}{2}-\beta\right)}{n}\right) \\
\zeta &= \frac{3}{8}+\frac{\beta}{4} && = \frac{1}{2}-\frac{1}{4}\left(\frac{1}{2}-\beta\right) \\
\sigma &= 64 \cdot \left\lceil \frac{6\pi n}{\left(\frac{1}{2}-\beta\right)^2} \right\rceil \cdot \left\lceil \log_e\left(\frac{32}{\delta}\right) \right\rceil && \le O\left(\frac{n}{\left(\frac{1}{2}-\beta\right)^2}\log\left(\frac{1}{\delta}\right)\right) \\
\ell &= \left\lceil \frac{6\pi n}{\left(\frac{1}{2}-\beta\right)^2} \right\rceil \cdot \sigma && \le O\left(\frac{n^2}{\left(\frac{1}{2}-\beta\right)^4}\log\left(\frac{1}{\delta}\right)\right)
\end{aligned}$$

Let $s_i^0 = 0$ for every $i \in [N]$.
For $j = 1, \ldots, \ell$:
  Draw $p^j \sim \overline{D_{\alpha,\zeta}}$ and $c^j_{1\cdots N} \sim p^j$.
  Issue $c^j \in \{\pm 1\}^N$ as a challenge and receive $a^j \in \{\pm 1\}$ as the response.
  For $i \in [N]$, let $s_i^j = s_i^{j-1} + a^j \cdot \phi^{p^j}(c_i^j)$.
  Accuse $I^j = \left\{ i \in [N] \mid s_i^j > \sigma \right\}$.

Figure 2: The interactive fingerprinting code $\mathcal{F} = \mathcal{F}_{n,N,\delta,\beta}$

2.3 Analysis Overview

Intuitively, the quantity s_i^j, which we call the *score* of user i, measures the "correlation" between the answers $(a^1, \cdots, a^j)$ of $\mathcal{P}$ and the i-th codeword $(c_i^1, \cdots, c_i^j)$, using a particular measure of correlation that takes into account the choices $p^1, \ldots, p^j$. If s_i^j ever exceeds the threshold σ, meaning that the answers are significantly correlated with the i-th codeword, then we accuse user i. Thus, our goal is to show two things: *Soundness*, that the score of an *innocent* user (i.e. $i \notin S^1$) never exceeds the threshold, as the answers cannot be correlated with the unknown i-th codeword. And *completeness*, that the score of every *guilty* user (i.e. $i \in S^1$) will at some point exceed the threshold, meaning that the answers must correlate with the i-th codeword for every $i \in S^1$.

2.3.1 Soundness

The proof of soundness closely mirrors Tardos' analysis [Tar08] of the non-interactive case. If i is innocent, then, since $\mathcal{P}$ doesn't see the codeword $(c_i^1, \cdots, c_i^j)$ of the i^{th} user, there cannot be too much correlation. In this case, one can show that s_i^j is the sum of j independent random variables, each with mean 0 and variance 1, where we take the answers $a^1, \ldots, a^j$ as fixed and the randomness is over the choice of the unknown codeword. By analogy to Gaussian random

variables, one would expect that $s_i^j \leq \sigma = \Theta(\sqrt{\ell \log(1/\delta)})$ with probability at least $1-\delta$. Formally, the fact that the score in each round is not bounded prevents the use of a Chernoff bound. But nonetheless, in Section 2.4, we prove soundness using a Chernoff-like tail bound for s_i^j.

2.3.2 Completeness

To prove completeness, we must show that, for guilty users $i \in S^1$, we have $s_i^j > \sigma$ for some $j \in [\ell]$ with high probability. In Sections 2.5.1 and 2.5.3, we prove that if $\mathcal{P}$ gives consistent answers in a $1-\beta$ fraction of rounds, then the sum of the scores for each of the guilty users is large. Specifically, in Theorem 2.15, we prove that with high probability

$$\sum_{i \in S^1} s_i^\ell \geq \Theta(\ell) \tag{1}$$

The constants hidden by the asymptotic notation are set to imply that, for at least one $i \in S^1$, the score s_i^ℓ is above the threshold $\sigma = \Theta(\ell/n)$. This step is not too different from the analysis of Tardos and Bun et al. [Tar08, BUV14] for the non-interactive case. To show that, for *every* $i \in S^1$, we will have $s_i^j > \sigma$ at some point, we must depart from the analysis of non-interactive fingerprinting codes. If $s_i^j > \sigma$, and user i is accused in round j, then the adversary will not see the suffix of codeword $(c_i^{j+1}, \cdots, c_i^\ell)$. By the same argument that was used to prove soundness, the answers will not be correlated with this suffix, so with high probability the score s_i^ℓ does not increase much beyond σ. Thus,

$$\sum_{i \in S^1} s_i^\ell \leq n \cdot O(\sigma) = \Theta(\ell). \tag{2}$$

The hidden constants are set to ensure that Equation (2) conflicts with Equation (1). Thus, we can conclude that $\mathcal{P}$ cannot give consistent answers for a $1-\beta$ fraction of rounds. That is to say, $\mathcal{P}$ is forced to be inconsistent because all of S^1 is accused and eventually $\mathcal{P}$ sees none of the codewords and is reduced to guessing an answer a^j.

2.3.3 Establishing Correlation

Proving Equation (1) is key to the analysis. Our proof thereof combines and simplifies the analyses of [Tar08] and [BUV14]. For this high level overview, we ignore the issue of robustness and fix $\beta = 0$.

First we prove that the correlation bound holds in expectation and then we show that it holds with high probability using an Azuma-like concentration bound. (Again, as the random variables being summed are not bounded, we are forced to use a more tailored analysis to prove concentration.) We show that it holds in expectation for each round. In Proposition 2.12, we show that the concentration grows in expectation in each round. For every $j \in [\ell]$,

$$\mathbb{E}\left[\sum_{i \in S^1} s_i^j - s_i^{j-1}\right] = \mathbb{E}\left[\sum_{i \in S^1} a^j \cdot \phi^{p^j}(c_i^j)\right] \geq \Omega(1), \tag{3}$$

where the expectations are taken over the randomness of p^j, c^j, and a^j. Equation (3), combined with a concentration result, implies Equation (1).

The intuition behind Equation (3) and the choice of p^j is as follows. Consistency guarantees that, if $c_i^j = b$ for all $i \in S^1$, then $a^j = b$. This is a weak correlation guarantee, but it suffices to ensure correlation between a^j and $\sum_{i \in S^1} c_i^j$. The affine scaling ϕ^{p^j} ensures that $\phi^{p^j}(c_i^j)$ has mean zero (i.e. is uncorrelated with a constant) and and unit variance (i.e. has unit correlation with itself). The expectation of $a^j \cdot \phi^{p^j}(c_i^j)$ can be interpreted as the i-th first-order Fourier coefficient of a^j as a function of c^j. To understand first-order Fourier coefficients, consider the "dictator" function: Suppose $a^j = c_{i^*}^j$ for some $i^* \in S^1$ - that is, $\mathcal{P}$ always outputs the i^*-th bit. Then

$$\mathop{\mathbb{E}}_{a^j, c^j, p^j}\left[a^j \sum_{i \in S^1} \phi^{p^j}(c_i^j)\right] = \mathop{\mathbb{E}}_{c^j, p^j}\left[c_{i^*}^j \cdot \phi^{p^j}(c_{i^*}^j)\right] = \mathop{\mathbb{E}}_{p^j}\left[2\sqrt{p^j(1-p^j)}\right] = \Theta(1).$$

This example can be generalised to a^j being an arbitrary function of $c_{S^1}^j$ using Fourier analysis. This calculation also indicates why we choose the probability density function of $p^j \sim D_{\alpha,1-\alpha}$ to be proportional to $1/\sqrt{p(1-p)}$.

To handle robustness ($\beta > 0$) we use the ideas of [BUV14]. With probability 2ζ each round is a "special" constant round—i.e. $c^j = (1)^N$ or $c^j = (-1)^N$. Otherwise it is a "normal" round where c^j is sampled as before. Intuitively, the adversary $\mathcal{P}$ cannot distinguish the special rounds from the normal rounds in which c happens to be constant. If the adversary gives inconsistent answers on normal rounds, then it must also give inconsistent answers on special rounds. Since there are many more special rounds than normal rounds, this means that a small number of inconsistencies in normal rounds implies a large number of inconsistencies on special rounds. Conversely, inconsistencies are absorbed by the special rounds, so we can assume there are very few inconsistencies in normal rounds. Thus $\mathcal{P}$ is forced to behave consistently on the normal rounds and the analysis on these rounds proceeds as before.

2.4 Proof of Soundness

We first show that no user is falsely accused except with probability $\delta/2$. This boils down to proving a concentration bound. Then another concentration bound shows that with high probability at most a δ fraction of users are falsely accused.

These concentrations bounds are essentially standard. However, we are showing concentration of sums of variables of the form $\phi^p(c)$, which may be quite large if $p \approx 0$ or $p \approx 1$. This technical problem prevents us from directly applying standard concentration bounds. Instead we open up the standard proofs and verify the desired concentration. We take the usual approach of bounding the moment generating function and using that to give a tail bound.

Lemma 2.3. *For $p \in [\alpha, 1-\alpha] \cup \{0,1\}$ and $t \in [-\sqrt{\alpha}/2, \sqrt{\alpha}/2]$, we have*

$$\mathop{\mathbb{E}}_{c \sim p}\left[e^{t\phi^p(c)}\right] \leq e^{t^2}.$$

Proof. If $p \in \{0,1\}$, $\phi^p = 0$ and the result is trivial. We have $\mathop{\mathbb{E}}_{c \sim p}[\phi^p(c)] = 0$, $\mathop{\mathbb{E}}_{c \sim p}\left[\phi^p(c)^2\right] = 1$, and, for $c \in \{\pm 1\}$, $|\phi^p(c)| \leq 1/\sqrt{\alpha}$. In particular, $|\phi^p(c) \cdot t| \leq 1/2$. For $u \in [-1/2, 1/2]$, we have $e^u \leq 1 + u + u^2$. Thus

$$\mathop{\mathbb{E}}_{c \sim p}\left[e^{t\phi^p(c)}\right] \leq 1 + t \mathop{\mathbb{E}}_{c \sim p}[\phi^p(c)] + t^2 \mathop{\mathbb{E}}_{c \sim p}\left[\phi^p(c)^2\right] = 1 + t^2 \leq e^{t^2}.$$

□

Lemma 2.4. *Let $p_1 \cdots p_m \in [\alpha, 1-\alpha] \cup \{0,1\}$ and $c_1 \cdots c_m$ drawn independently with $c_i \sim p_i$. Let $a_1 \cdots a_m \in [-1,1]$ be fixed. For all $\lambda \geq 0$, we have*

$$\mathbb{P}\left[\sum_{i\in[m]} a_i \phi^{p_i}(c_i) \geq \lambda\right] \leq e^{-\lambda^2/4m} + e^{-\sqrt{\alpha}\lambda/4}.$$

Proof. By Lemma 2.3, for all $t \in [-\sqrt{\alpha}/2, \sqrt{\alpha}/2]$,

$$\mathop{\mathbb{E}}_{c}\left[e^{t\sum_{i\in[m]} a_i \phi^{p_i}(c_i)}\right] \leq \prod_{i\in[m]} \mathop{\mathbb{E}}_{c_i}\left[e^{t a_i \phi^{p_i}(c_i)}\right] \leq e^{t^2 m}.$$

By Markov's inequality,

$$\mathbb{P}\left[\sum_{i\in[m]} a_i \phi^{p_i}(c_i) \geq \lambda\right] \leq \frac{\mathbb{E}\left[e^{t\sum_{i\in[m]} a_i \phi^{p_i}(c_i)}\right]}{e^{t\lambda}} \leq e^{t^2 m - t\lambda}.$$

Set $t = \min\{\sqrt{\alpha}/2, \lambda/2m\}$. If $\lambda \in [0, m\sqrt{\alpha}]$, then

$$\mathbb{P}\left[\sum_{i\in[m]} a_i \phi^{p_i}(c_i) \geq \lambda\right] \leq e^{-\lambda^2/4m}.$$

On the other hand, if $\lambda \geq m\sqrt{\alpha}$, then

$$\mathbb{P}\left[\sum_{i\in[m]} a_i \phi^{p_i}(c_i) \geq \lambda\right] \leq e^{\alpha m/4 - \sqrt{\alpha}\lambda/2} \leq e^{-\sqrt{\alpha}\lambda/4}.$$

The result is obtained by adding these expressions. □

The following theorem shows how we can beat the union bound for tail bounds on partial sums.

Theorem 2.5 (Etemadi's Inequality [Ete85])**.** *Let $X_1 \cdots X_n \in \mathbb{R}$ be independent random variables. For $k \in [n]$, define $S_k = \sum_{i\in[k]} X_i$ to be the k^{th} partial sum. Then, for all $\lambda > 0$,*

$$\mathbb{P}\left[\max_{k\in[n]} |S_k| > 4\lambda\right] \leq 4 \cdot \max_{k\in[n]} \mathbb{P}\left[|S_k| > \lambda\right].$$

Proposition 2.6 (Individual Soundness)**.** *For all $i \in [N]$, we have*

$$\mathbb{P}\left[i \in I \setminus S^1\right] \leq 8(e^{-\sigma^2/64\ell} + e^{-\sigma\sqrt{\alpha}/16}) \leq \delta/2,$$

where the probability is taken over $\mathsf{IFPC}_{N,N,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ for an arbitrary $\mathcal{P}$.

Proof. Let $i \in [N] \setminus S^1$. Since the adversary does not see c_i^j for any $j \in [\ell]$, we may treat the answers of the adversary as fixed and analyse s_i^j as if c_i^j was drawn after the actions of the adversary are fixed. Thus, by Lemma 2.4, for every $j \in [\ell]$,

$$\mathbb{P}\left[s_i^j > \frac{\sigma}{4}\right] = \mathbb{P}\left[\sum_{k\in[j]} a^k \phi^{p^k}(c_i^k) > \frac{\sigma}{4}\right] \leq e^{-\sigma^2/64\ell} + e^{-\sigma\sqrt{\alpha}/16}.$$

Likewise $\mathbb{P}\left[s_i^j < -\frac{\sigma}{4}\right] \le e^{-\sigma^2/64\ell} + e^{-\sigma\sqrt{\alpha}/16}$. Thus, by Theorem 2.5,

$$\mathbb{P}[i \in I] \le \mathbb{P}\left[\max_{j\in[\ell]} |s_i^j| > \sigma\right] \le 4 \max_{j\in[\ell]} \mathbb{P}\left[|s_i^j| > \frac{\sigma}{4}\right] \le 8(e^{-\sigma^2/64\ell} + e^{-\sigma\sqrt{\alpha}/16}) \le \frac{\delta}{2}.$$

□

Theorem 2.7 (Soundness)**.** *We have*

$$\mathbb{P}\left[|I \setminus S^1| > \delta N\right] \le \min\left\{\delta N, e^{-\delta N/8}\right\},$$

where the probability is taken over $\mathsf{IFPC}_{N,N,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ *for an arbiratry* $\mathcal{P}$*.*

Interestingly, this theorem does not require $|S^1| \le n$ – that is, it holds with respect to $\mathsf{IFPC}_{N,N,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$, rather than $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$. It only requires that $\mathcal{F}$ does not see the codewords of users not in S^1.

Proof. Let $E_i \in \{0,1\}$ be the indicator of the event $i \in I \setminus S^1$. The E_is for $i \in [N]$ are independent (conditioned on the choice of S^1 and p^j for $j \in [\ell]$). Moreover, by Proposition 2.6, $\mathbb{E}[E_i] \le \delta/2$ for all $i \in [N] \setminus S^1$. Thus, by a Chernoff bound,

$$\mathbb{P}\left[|I \setminus S^1| > \delta N\right] = \mathbb{P}\left[\sum_{i\in[N]} E_i > \delta N\right] \le e^{-\delta N/8}.$$

If $\delta < 1/N$, then this is a very poor bound. Instead we use the fact that the E_is are discrete and Markov's inequality, which amounts to a union bound. For $\delta N < 1$, we have

$$\mathbb{P}\left[|I \setminus S^1| > \delta N\right] = \mathbb{P}\left[|I \setminus S^1| \ge 1\right] \le \mathbb{E}\left[\sum_{i\in[N]} E_i\right] \le \frac{\delta N}{2} \le \delta N.$$

□

The following lemma will be useful later.

Lemma 2.8. *For $i \in [N]$, let $j_i \in [\ell+1]$ be the first j such that $i \notin S^j$, where we define $S^{\ell+1} = \emptyset$. For any $S \subset [N]$,*

$$\mathbb{P}\left[\sum_{i\in S} s_i^\ell - s_i^{j_i-1} > \lambda\right] \le e^{-\lambda^2/4|S|\ell} + e^{-\sqrt{\alpha}\lambda/4},$$

where the probability is taken over $\mathsf{IFPC}_{N,N,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ *for an arbitrary* $\mathcal{P}$*.*

Proof. We have

$$\sum_{i\in S} s_i^\ell - s_i^{j_i-1} = \sum_{i\in S} \sum_{j\in[\ell]} \mathbb{I}(j \ge j_i) a^j \phi^{p^j}(c_i^j).$$

Again, since the adversary doesn't see c_i^j for $j \ge j_i$, the random variables $\mathbb{I}(j \ge j_i)a^j$ and $\phi^{p^j}(c_i^j)$ are independent, so we can view $\mathbb{I}(j \ge j_i)a^j \in [-1,1]$ as fixed. Now the result follows from Lemma 2.4. □

2.5 Proof of Completeness

To show that the fingerprinting code identifies guilty users we must lower bound the scores $\sum_{i \in S^1} s_i^\ell$. First we bound their expectation and then their tails.

2.5.1 Biased Fourier Analysis

For this section, assume that the adversary $\mathcal{P}$ is always consistent - that is, we have no robustness and $\beta = 0$. Robustness will be added in Section 2.5.2. Here we establish that the scores have good expectation, namely

$$\mathbb{E}\left[\sum_{i \in S^1} s_i^j - s_i^{j-1}\right] \geq \Omega(1)$$

for all $j \in [\ell]$. The score s_i^ℓ computes the 'correlation' between the bits given to user i and the output of the adversary. We must show that that the adversary's consistency constraint implies that there exists some correlation on average.

In this section we deviate from the proof in [Tar08]. We use biased Fourier analysis to give a more intuitive proof of the correlation bound.

We have the following lemma and proposition, which relate the correlation $a^j \cdot \sum_{i \in S^1} \phi^{p^j}(c_i^j)$ to the properties of a^j as a function of p^j. To interpret these imagine that f represents the adversary $\mathcal{P}$ with one round viewed in isolation – the fingerprinting code gives the adversary c^j and the adversary responds with $f(c^j_{S^j})$.

Firstly, the following lemma gives an interpretation of the correlation value for a fixed p^j.

Lemma 2.9. *Let $f : \{\pm 1\}^n \to \mathbb{R}$. Define $g : [0,1] \to \mathbb{R}$ by $g(p) = \mathop{\mathbb{E}}_{c_{1\cdots n} \sim p}[f(c)]$. For any $p \in (0,1)$,*

$$\mathop{\mathbb{E}}_{c_{1\cdots n} \sim p}\left[f(c) \cdot \sum_{i \in [n]} \phi^p(c_i)\right] = g'(p)\sqrt{p(1-p)}.$$

Proof. For $p \in (0,1)$ and $s \subset [n]$, define $\phi_s^p : \{\pm 1\}^n \to \mathbb{R}$ by $\phi_s^p(c) = \prod_{i \in s} \phi^p(c_i)$. The functions ϕ_s^p form an orthonormal basis with respect to the product distribution with bias p – that is,

$$\forall s, t \subset [n] \quad \mathop{\mathbb{E}}_{c_{1\cdots n} \sim p}\left[\phi_s^p(c) \cdot \phi_t^p(c)\right] = \left\{ \begin{array}{ll} 1 & s = t \\ 0 & s \neq t \end{array} \right\}.$$

Thus, for any $p \in (0,1)$, we can write f in terms of these basis functions:

$$\forall c \in \{\pm 1\}^n \quad f(c) = \sum_{s \subset [n]} \tilde{f}^p(s) \phi_s^p(c),$$

where

$$\forall s \subset [n] \quad \tilde{f}^p(s) = \mathop{\mathbb{E}}_{c_{1\cdots n} \sim p}\left[f(c)\phi_s^p(c)\right].$$

This decomposition is a generalisation of Fourier analysis to biased distributions [O'D14, §8.4].

For $p, q \in (0,1)$, the expansion of f gives the following expressions for $g(q)$, $g'(q)$ and $g'(p)$.

$$\begin{aligned}
g(q) &= \mathop{\mathbb{E}}_{c_{1\cdots n}\sim q}[f(c)] \\
&= \sum_{s\subset[n]} \tilde{f}^p(s) \mathop{\mathbb{E}}_{c_{1\cdots n}\sim q}\left[\phi_s^p(c)\right] \\
&= \sum_{s\subset[n]} \tilde{f}^p(s) \prod_{i\in s} \mathop{\mathbb{E}}_{c\sim q}[\phi^p(c)] \\
&= \sum_{s\subset[n]} \tilde{f}^p(s) \left(q\sqrt{\frac{1-p}{p}} - (1-q)\sqrt{\frac{p}{1-p}}\right)^{|s|}. \\
g'(q) &= \sum_{s\subset[n]:s\neq\emptyset} \tilde{f}^p(s)\cdot|s|\cdot\left(q\sqrt{\frac{1-p}{p}} - (1-q)\sqrt{\frac{p}{1-p}}\right)^{|s|-1}\cdot\left(\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}}\right). \\
g'(p) &= \sum_{s\subset[n]:s\neq\emptyset} \tilde{f}^p(s)\cdot|s|\cdot 0^{|s|-1}\cdot\left(\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}}\right) \\
&= \sum_{i\in[n]} \tilde{f}^p(\{i\})\cdot\left(\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}}\right).
\end{aligned}$$

Note that $\tilde{f}^p(\{i\}) = \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)\phi^p(c_i)]$ and, hence,

$$\mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}\left[f(c)\cdot\sum_{i\in[n]}\phi^p(c_i)\right] = \sum_{i\in[n]}\tilde{f}^p(\{i\}) = \frac{g'(p)}{\sqrt{\frac{1-p}{p}}+\sqrt{\frac{p}{1-p}}} = g'(p)\sqrt{p(1-p)}.$$

□

Now we can interpret the correlation for a random $p^j \sim D_{a,b}$.

Proposition 2.10. *Let $f : \{\pm1\}^n \to \mathbb{R}$. Define $g : [0,1] \to \mathbb{R}$ by $g(p) = \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)]$. For any $0 \le a < b \le 1$,*

$$\mathop{\mathbb{E}}_{p\sim D_{a,b}}\left[\mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}\left[f(c)\cdot\sum_{i\in[n]}\phi^p(c_i)\right]\right] = \frac{g(b)-g(a)}{2\sin^{-1}(\sqrt{b})-2\sin^{-1}(\sqrt{a})} \ge \frac{g(b)-g(a)}{\pi}.$$

This effectively follows by integrating Lemma 2.9.

Proof. Let $\mu(p) = C_{a,b}/\sqrt{p(1-p)}$ be the probability density function for the distribution $D_{a,b}$ on the interval (a,b). By Lemma 2.9 and the fundamental theorem of calculus, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{p\sim D_{a,b}}\left[\mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}\left[f(c)\cdot\sum_{i\in[n]}\phi^p(c_i)\right]\right] &= \mathop{\mathbb{E}}_{p\sim D_{a,b}}\left[g'(p)\sqrt{p(1-p)}\right] \\
&= \int_a^b g'(p)\sqrt{p(1-p)}\mu(p)\mathrm{d}p \\
&= C_{a,b}\int_a^b g'(p)\mathrm{d}p \\
&= C_{a,b}\cdot(g(b)-g(a)).
\end{aligned}$$

It remains to show that $C_{a,b} = \left(2\sin^{-1}(\sqrt{b}) - 2\sin^{-1}(\sqrt{a})\right)^{-1} \geq 1/\pi$. This follows from observing that

$$C_{a,b}^{-1} = \int_a^b \frac{1}{\sqrt{p(1-p)}}\mathrm{d}p = \int_a^b \left(\frac{\mathrm{d}}{\mathrm{d}p} 2\sin^{-1}(\sqrt{p})\right)\mathrm{d}p = 2\sin^{-1}(\sqrt{b}) - 2\sin^{-1}(\sqrt{a})$$

and

$$C_{a,b}^{-1} \leq C_{0,1}^{-1} = 2\sin^{-1}(1) - 2\sin^{-1}(0) = \pi.$$

□

Now we have a lemma to bring consistency into the picture. If f is consistent, $b \approx 1$, and $a \approx 0$, then

$$g(b) - g(a) \approx g(1) - g(0) = f((1)^n) - f((-1)^n) = 1 - (-1) = 2.$$

This gives a lower bound on the correlation for consistent f.

Lemma 2.11. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$. Define $g : [0,1] \to [-1,1]$ by $g(p) = \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)]$. Suppose $\alpha \in [0,1]$. Then $|g(1-\alpha) - g(1)| \leq 2n\alpha$ and $|g(\alpha) - g(0)| \leq 2n\alpha$.*

Proof. We have $\mathop{\mathbb{P}}_{c_{1\cdots n}\sim 1-\alpha}[X = (1)^n] = (1-\alpha)^n$ and

$$\begin{aligned} g(1-\alpha) - g(1) =& f((1)^n) \cdot \mathop{\mathbb{P}}_{c_{1\cdots n}\sim 1-\alpha}[c = (1)^n] + \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c) | c \neq (1)^n] \cdot \mathop{\mathbb{P}}_{c_{1\cdots n}\sim 1-\alpha}[c \neq (1)^n] - g(1) \\ =& g(1) \cdot (1-\alpha)^n + \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c) | c \neq (1)^n] \cdot (1 - (1-\alpha)^n) - g(1) \\ =& \left(g(1) - \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c) | c \neq (1)^n]\right) \cdot ((1-\alpha)^n - 1). \end{aligned}$$

Now $\left|g(1) - \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c) | c \neq (1)^n]\right| \leq 2$ and $|(1-\alpha)^n - 1| \leq n\alpha$, whence $|g(1-\alpha) - g(1)| \leq 2n\alpha$. The other half of the lemma is symmetric.

□

2.5.2 Robustness

We require the fingerprinting code to be robust to inconsistent answers. We show that the correlation is still good in the presence of inconsistencies.

For $f : \{\pm 1\}^n \to \{\pm 1\}$, define a random variable $\xi_{\alpha,\zeta}(f)$ by

$$\xi_{\alpha,\zeta}(f) = f(c) \cdot \sum_{i\in[n]} \phi^p(c_i) + \gamma \mathbb{I}(p \in \{0,1\} \wedge f(c) \neq 2p - 1), \quad p \sim \overline{D_{\alpha,\zeta}}, \quad c_{1\cdots n} \sim p,$$

where $\mathbb{I}$ is the indicator function and $\gamma \in (0, 1/2)$ satisfies $\zeta\gamma/2 = (1-2\zeta)/\pi$ - that is,

$$\gamma := \frac{2}{\pi}\frac{1-2\zeta}{\zeta}.$$

The first term $f(c) \cdot \sum_{i\in[n]} \phi^p(c_i)$ measures the correlation as before. The second term $\gamma\mathbb{I}(p \in \{0,1\} \wedge f(c) \neq 2p-1)$ measures inconsistencies. We will lower bound the expectation of $\xi_{\alpha,\zeta}(f)$, which amounts to saying "either there is good correlation *or* there is an inconsistency with good probability." Thus either the fingerprinting code is able to accuse users or the adversary is forced to be inconsistent.

The following bounds the expected increase in scores from one round of interaction.

Proposition 2.12. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ and $\alpha, \zeta \in (0, 1/2)$. Then*

$$\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] \geq \frac{2}{\pi}(1-2\zeta)(1-2n\alpha).$$

Proof. Define $g : [0,1] \to [-1,1]$ by $g(p) = \mathbb{E}_{c_{1\cdots n}\sim p}[f(c)]$. Now

$$\begin{aligned}
\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] =& \mathbb{P}_{p\sim\overline{D_{\alpha,\zeta}}}[p=0]\cdot\gamma\mathbb{I}(f((-1)^n)=1) + \mathbb{P}_{p\sim\overline{D_{\alpha,\zeta}}}[p=1]\cdot\gamma\mathbb{I}(f((1)^n)=-1) \\
&+ \mathbb{P}_{p\sim\overline{D_{\alpha,\zeta}}}[p\in[\alpha,1-\alpha]]\cdot\mathbb{E}_{p\sim D_{\alpha,1-\alpha}}\left[\mathbb{E}_{c_{1\cdots n}\sim p}\left[f(c)\cdot\sum_{i\in[n]}\phi^p(c_i)\right]\right] \\
=& \zeta\cdot\gamma\left(\mathbb{I}(g(0)=1)+\mathbb{I}(g(1)=-1)\right) \\
\text{(by Proposition 2.10)}\quad &+ (1-2\zeta)\cdot\frac{g(1-\alpha)-g(\alpha)}{2\sin^{-1}(\sqrt{1-\alpha})-2\sin^{-1}(\sqrt{\alpha})} \\
\geq& \zeta\cdot\gamma\left(\frac{1+g(0)}{2}+\frac{1-g(1)}{2}\right)+(1-2\zeta)\cdot\frac{g(1-\alpha)-g(\alpha)}{\pi} \\
=& \frac{1-2\zeta}{\pi}\left(1+g(0)+1-g(1)+g(1-\alpha)-g(\alpha)\right) \\
\geq& \frac{1-2\zeta}{\pi}\left(2-|g(\alpha)-g(0)|-|g(1-\alpha)-g(1)|\right) \\
\text{(by Lemma 2.11)}\quad \geq& \frac{1-2\zeta}{\pi}(2-4n\alpha).
\end{aligned}$$

□

2.5.3 Concentration

So far we have shown that the fingerprinting code achieves good correlation or the adversary is not consistent *in expectation*. However, we need this to hold with high probability. Thus we now show that sums of $\xi_{\alpha,\zeta}(f)$ variables concentrate around their expectation.

Again, the proofs in this section are standard. However, the $\xi_{\alpha,\zeta}(f)$ variables can be quite unwieldy and we are thus unable to apply standard results directly. So instead we must open the proofs and verify that the concentration bounds hold. We proceed by bounding the moment generating function of $\xi_{\alpha,\zeta}(f)$ and then proving an Azuma-like concentration inequality. These calculations are not novel or insightful.

Proposition 2.13. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$, $\alpha \in (0, 1/2)$, $\zeta \in [1/4, 1/2)$, and $t \in [-\sqrt{\alpha}/8, \sqrt{\alpha}/8]$. Then*

$$\mathbb{E}\left[e^{t(\xi_{\alpha,\zeta}(f)-\mathbb{E}[\xi_{\alpha,\zeta}(f)])}\right] \leq e^{Ct^2},$$

where $C = \frac{64e^{n\alpha/4}}{\alpha}$.

Proof. We have

$$\xi_{\alpha,\zeta}(f) = f(c)\cdot\sum_{i\in[n]}\phi^p(c_i) + \gamma\mathbb{I}\left(p\in\{0,1\}\wedge f(c)\neq 2p-1\right), \quad p\sim\overline{D_{\alpha,\zeta}}, \quad c_{1\cdots n}\sim p.$$

Let $Y = \sum_{i\in[n]} \phi^p(c_i)$. By Lemma 2.3 and independence,

$$\mathbb{E}\left[e^{tY}\right] = \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}\left[e^{t\sum_{i\in[n]}\phi^p(c_i)}\right] = \left(\mathop{\mathbb{E}}_{c\sim p}\left[e^{t\phi^p(c)}\right]\right)^n \le e^{t^2 n}$$

for $t \in [-\sqrt{\alpha}/2, \sqrt{\alpha}/2]$. Pick $t \in \{\pm\sqrt{\alpha}/2\}$ such that

$$\sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!}\mathbb{E}\left[Y^{2k+1}\right] \ge 0.$$

Then by dropping positive terms, for all $j \ge 1$,

$$0 \le \mathbb{E}\left[Y^{2j}\right] \le \frac{(2j)!}{t^{2j}}\sum_{k=0}^{\infty}\frac{t^k}{k!}\mathbb{E}\left[Y^k\right] = \frac{(2j)!}{t^{2j}}\mathbb{E}\left[e^{tY}\right] \le \frac{(2j)!}{t^{2j}}e^{nt^2} = \frac{4^j(2j)!}{\alpha^j}e^{n\alpha/4}.$$

Thus we have bounded the even moments of Y. By Cauchy-Schwartz, for $k = 2j+1 \ge 3$,

$$\mathbb{E}\left[|Y|^k\right] \le \sqrt{\mathbb{E}\left[Y^{2j}\right]\cdot\mathbb{E}\left[Y^{2j+2}\right]} \le \sqrt{\frac{4^j(2j)!}{\alpha^j}e^{n\alpha/4}\cdot\frac{4^{j+1}(2j+2)!}{\alpha^{j+1}}e^{n\alpha/4}} = \frac{2^k k!}{\alpha^{k/2}}e^{n\alpha/4}\sqrt{\frac{k+1}{k}}.$$

Since $|f(c)| \le 1$, we have $\mathbb{E}\left[|f(c)\cdot Y|^k\right] \le \mathbb{E}\left[|Y|^k\right] \le 2^{k+1}k!e^{n\alpha/4}/\alpha^{k/2}$ for all $k \ge 2$. Since $\zeta \in [1/4, 1/2)$, we have $\gamma = (2/\pi)(1-2\zeta)/\zeta \in (0,1)$. Hence $\mathbb{E}\left[|\gamma\mathbb{I}(p \in \{0,1\} \wedge f(c) \ne 2p-1)|^k\right] \le 1$ for all k. The map $u \mapsto |u|^k$ is convex for all $k \ge 2$, thus $|(x+y)/2|^k \le (|x|^k + |y|^k)/2$ for all $k \ge 2$ and $x, y \in \mathbb{R}$. Combining these three facts, we have

$$\mathbb{E}\left[|\xi_{\alpha,\zeta}(f)|^k\right] \le 2^{k-1}\mathbb{E}\left[|f(c)\cdot Y|^k + |\gamma\mathbb{I}(f(c) \ne f^*(c))|^k\right] \le \frac{2^{2k}k!e^{n\alpha/4}}{\alpha^{k/2}} + 2^{k-1} \le \frac{2^{2k+1}k!e^{n\alpha/4}}{\alpha^{k/2}}.$$

For $t \in [-\sqrt{\alpha}/8, \sqrt{\alpha}/8]$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{t\xi_{\alpha,\zeta}(f)}\right] &\le 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + \sum_{k=2}^{\infty}\frac{|t|^k}{k!}\mathbb{E}\left[|\xi_{\alpha,\zeta}(f)|^k\right] \\
&\le 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + \sum_{k=2}^{\infty}\frac{|t|^k}{k!}\frac{2^{2k+1}k!e^{n\alpha/4}}{\alpha^{k/2}} \\
&= 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + 2e^{n\alpha/4}\sum_{k=2}^{\infty}\left(\frac{4|t|}{\sqrt{\alpha}}\right)^k \\
&\le 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + 2e^{n\alpha/4}\sum_{k=2}^{\infty}\left(\frac{4|t|}{\sqrt{\alpha}}\right)^2 2^{-(k-2)} \\
&= 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + \frac{64e^{n\alpha/4}}{\alpha}t^2 \\
&\le e^{t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + Ct^2}
\end{aligned}$$

□

Theorem 2.14 (Azuma-Doob Inequality). *Let $X_1 \cdots X_m \in \mathbb{R}$, $\mu_1 \cdots \mu_m \mathbb{R}$ and $\mathcal{U}_0 \cdots \mathcal{U}_m \in \Omega$ be random variables such that, for all $i \in [m]$,*

- *X_i is determined by $\mathcal{U}_i$,*
- *μ_i is determined by $\mathcal{U}_{i-1}$, and*
- *$\mathcal{U}_{i-1}$ is determined by $\mathcal{U}_i$.*

Suppose that, for all $i \in [m]$, $u \in \Omega$, and $t \in [-c, c]$,

$$\mathbb{E}\left[e^{t(X_i - \mu_i)} \mid \mathcal{U}_{i-1} = u\right] \leq e^{Ct^2}.$$

If $\lambda \in [0, 2Cmc]$, then

$$\mathbb{P}\left[\left|\sum_{i\in[m]} (X_i - \mu_i)\right| \geq \lambda\right] \leq 2e^{-\lambda^2/4Cm}.$$

If $\lambda \geq 2Cmc$, then

$$\mathbb{P}\left[\left|\sum_{i\in[m]} (X_i - \mu_i)\right| \geq \lambda\right] \leq 2e^{mCc^2 - c\lambda} \leq 2e^{-c\lambda/2}.$$

Proof. First we show by induction on $k \in [m]$ that, for all $u \in \Omega$ and $t \in [-c, c]$,

$$\mathbb{E}\left[e^{t\sum_{i=m-k+1}^{m}(X_i - \mu_i)} \mid \mathcal{U}_{m-k} = u\right] \leq e^{k\cdot Ct^2}.$$

This clearly holds for $k = 1$, as this is our supposition for $i = m$. Now suppose this holds for some $k \in [m-1]$. For $u \in \Omega$ and $t \in [-c, c]$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{t\sum_{i=m-k}^{m}(X_i - \mu_i)} \mid \mathcal{U}_{m-(k+1)} = u\right] &= \sum_{v\in\Omega} \mathbb{P}\left[\mathcal{U}_{m-k} = v \mid \mathcal{U}_{m-k-1} = u\right] \mathbb{E}\left[e^{t\sum_{i=m-k}^{m}(X_i - \mu_i)} \mid \mathcal{U}_{m-k} = v\right] \\
&= \sum_{v\in\Omega} \mathbb{P}\left[v \mid u\right] \mathbb{E}\left[e^{t(X_{m-k} - \mu_{m-k})} e^{t\sum_{i=m-k+1}^{m}(X_i - \mu_i)} \mid v\right] \\
&\quad \text{(using shorthand } v \equiv \mathcal{U}_{m-k} = v \text{ and } u \equiv \mathcal{U}_{m-k-1} = u\text{)} \\
&= \sum_{v\in\Omega} \mathbb{P}\left[v \mid u\right] \mathbb{E}\left[e^{t(X_{m-k} - \mu_{m-k})} \mid v\right] \mathbb{E}\left[e^{t\sum_{i=m-k+1}^{m}(X_i - \mu_i)} \mid v\right] \\
&\quad \text{(since } U_{m-k} = v \text{ determines } X_{m-k} \text{ and } \mu_{m-k}\text{)} \\
&\leq \sum_{v\in\Omega} \mathbb{P}\left[v \mid u\right] \mathbb{E}\left[e^{t(X_{m-k} - \mu_{m-k})} \mid v\right] e^{k\cdot Ct^2} \\
&\quad \text{(by the induction hypothesis)} \\
&= \mathbb{E}\left[e^{t(X_{m-k} - \mu_{m-k})} \mid u\right] e^{k\cdot Ct^2} \\
&\leq e^{Ct^2} e^{k\cdot Ct^2} \\
&\quad \text{(by our supposition for } i = m - k\text{)} \\
&= e^{(k+1)\cdot Ct^2}.
\end{aligned}$$

Thus, for all $t \in [-c, c]$, we have

$$\mathbb{E}\left[e^{t\sum_{i=1}^{m}(X_i-\mu_i)}\right] \le e^{m\cdot Ct^2}.$$

By Markov's inequality we have

$$\mathbb{P}\left[\sum_{i\in[m]}(X_i-\mu_i) \ge \lambda\right] \le \frac{\mathbb{E}\left[e^{t\sum_{i\in[m]}(X_i-\mu_i)}\right]}{e^{t\lambda}} \le e^{mCt^2-t\lambda}$$

and

$$\mathbb{P}\left[\sum_{i\in[m]}(X_i-\mu_i) \le -\lambda\right] \le \frac{\mathbb{E}\left[e^{-t\sum_{i\in[m]}(X_i-\mu_i)}\right]}{e^{(-t)(-\lambda)}} \le e^{mCt^2-t\lambda}$$

for all $t \in [0, c]$ and $\lambda > 0$. Set $t = \min\{c, \lambda/2mC\}$ to obtain the result. □

2.5.4 Bounding the Score

Now we can finally show that the scores are large with high probability.

Theorem 2.15 (Correlation Lower Bound)**.** *At the end of* $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ *for arbitrary* $\mathcal{P}$*, we have, for any* $\lambda \in [0, 17.5\ell/\sqrt{\alpha}]$*,*

$$\gamma\theta^\ell + \sum_{i\in S^1} s_i^\ell \ge \frac{2}{\pi}(1-2\zeta)(1-2n\alpha)\ell - \lambda$$

with probability at least $1 - 2e^{-\frac{\lambda^2\alpha}{280\ell}}$*.*

Proof. Since the adversary $\mathcal{P}$ is computationally unbounded and arbitrary, we may assume it is deterministic. We may also assume $n = |S^1|$ and that the adversary is able to see $c^j_{S^1}$ at each round. (This only gives the adversary more power.)

This means that for each $j \in [\ell]$ we can define a function $f^j : \{\pm1\}^n \to \{\pm1\}$ that only depends on the interaction up to round $j-1$ (i.e. is a function of the state of $\mathcal{P}$ before it receives c^j) and satisfies $f^j(c^j_{S^j}) = a^j$. For $j \in [\ell]$, define

$$X_j := \gamma \cdot \mathbb{I}\left(p^j \in \{0,1\} \wedge f^j(c^j_{S^1}) \ne 2p^j - 1\right) + f^j(c^j_{S^1}) \cdot \sum_{i\in S^1} \phi^{p^j}(c^j_i) \sim \xi_{\alpha,\zeta}(f^j),$$

where $\sim$ denotes having the same distribution. We have

$$\gamma\cdot(\theta^j - \theta^{j-1}) + \sum_{i\in S^1}(s_i^j - s_i^{j-1}) \le X_j$$

and

$$\gamma\theta^\ell + \sum_{i\in S^1} s_i^\ell \le \sum_{j\in[\ell]} X_j \sim \sum_{j\in[\ell]} \xi_{\alpha,\zeta}(f^j).$$

Now we can apply the above lemmas to bound the expectation and tail of this random variable.

Firstly, Proposition 2.12 shows that

$$\mu_j := \mathbb{E}\left[X_j\right] = \mathbb{E}\left[\xi_{\alpha,\zeta}(f^j)\right] \ge \frac{2}{\pi}(1-2\zeta)(1-2n\alpha)$$

for all f^j. Moreover, by Proposition 2.13,

$$\mathbb{E}\left[e^{t(X^j-\mu_j)}\right] = \mathbb{E}\left[e^{t(\xi_{\alpha,\zeta}(f^j)-\mathbb{E}\left[\xi_{\alpha,\zeta}(f^j)\right])}\right] \le e^{Ct^2}$$

for all $t \in [-\sqrt{\alpha}/8, \sqrt{\alpha}/8]$, where $C = 70/\alpha \ge 64e^{n\alpha/4}/\alpha$, as $\alpha \le 1/4n$.

Define $\mathcal{U}_j = (f^1, p^1, c^1, \cdots, f^j, p^j, c^j, f^{j+1})$ for $j \in [\ell] \cup \{0\}$. Now $X_1 \cdots X_\ell$, $\mu_1 \cdots \mu_\ell$, and $\mathcal{U}_0, \cdots, \mathcal{U}_\ell$ satisfy the hypotheses of Theorem 2.14 with $C = 70/\alpha$, $c = \sqrt{\alpha}/8$, and $m = \ell$.

For $\lambda \in [0, 2Cmc] = [0, 17.5\ell/\sqrt{\alpha}]$, we have

$$\mathbb{P}\left[\sum_{j\in[\ell]} X_j \le \frac{2}{\pi}(1-2\zeta)(1-2n\alpha)\ell - \lambda\right] \le \mathbb{P}\left[\left|\sum_{i\in[m]}(X_i-\mu_i)\right| \ge \lambda\right] \le 2e^{-\lambda^2/4Cm} \le 2e^{-\frac{\lambda^2\alpha}{280\ell}},$$

as required. □

However, we can also prove that the scores are small with high probability. This follows from the fact that users with large scores are accused and therefore no user's score can be too large:

Lemma 2.16. *For all $\lambda > 0$,*

$$\mathbb{P}\left[\sum_{i\in S^1} s_i^\ell > \lambda + n\sigma + \frac{n}{\sqrt{\alpha}}\right] \le e^{-\lambda^2/4n\ell} + e^{-\sqrt{\alpha}\lambda/4},$$

where the probability is taken over $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ for an arbitrary $\mathcal{P}$.

We will set $\lambda = \sigma$ and, since $1/\sqrt{\alpha} \le \sigma$, we get that $\sum_{i\in S^1} s_i^\ell \le 3\sigma n$ with high probability.

Proof. Let $j_i \in [\ell+1]$ be as in Lemma 2.8 – that is, $i \notin S^{j_i}$ and $i \in S^{j_i-1}$, where we define $S^{\ell+1} = \emptyset$ and $S^0 = [N]$. By the definition of j_i, s^j, and S^j, we have $s_i^{j_i-2} \le \sigma$ for all $i \in S^1$, as otherwise $i \in I^{j_i-2}$ and therefore $i \notin S^{j_i-1} = S^{j_i-2} \setminus I^{j_i-2}$. If $i \in S^1$, then $j_i = 1$ and $s_i^{j_i-1} = 0$. Thus

$$\sum_{i\in S^1} s_i^{j_i-1} = \sum_{i\in S^1} s_i^{j_i-2} + a^{j_i-1}\phi^{p^{j_i-1}}(c_i^{j_i-1}) \le \sum_{i\in S^1} \sigma + \frac{1}{\sqrt{\alpha}} \le n\sigma + \frac{n}{\sqrt{\alpha}}.$$

By Lemma 2.8,

$$\mathbb{P}\left[\sum_{i\in S^1} s_i^\ell - s_i^{j_i-1} > \lambda\right] \le e^{-\lambda^2/4n\ell} + e^{-\sqrt{\alpha}\lambda/4}.$$

The lemma follows. □

Now we show that the conflicting bounds of Theorem 2.15 and Lemma 2.16 imply completeness - that is, the adversary $\mathcal{P}$ cannot be consistent.

Theorem 2.17 (Completeness)**.** *At the end of $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ for an arbitrary $\mathcal{P}$, we have $\theta^\ell > \beta\ell$ with probability at least $1 - \delta^{\frac{1}{2}(\frac{1}{2}-\beta)n}$, assuming $\left(\frac{1}{2}-\beta\right)n \ge 1$.*

Proof. Suppose for the sake of contradiction that $\theta^\ell \le \beta\ell$. By Lemma 2.16, $\sum_{i\in S^1} s_i^\ell \le \lambda + n\sigma + \frac{n}{\sqrt{\alpha}}$ with probability at least $1 - e^{-\lambda^2/4n\ell} - e^{-\sqrt{\alpha}\lambda/4}$. Set $\lambda = n\sigma \ge \frac{n}{\sqrt{\alpha}}$. Now we assume

$$\sum_{i\in S^1} s_i^\ell \le 3n\sigma,$$

which holds with probability at least $1 - e^{-n\sigma^2/4\ell} - e^{-\sqrt{\alpha}n\sigma/4}$. Then

$$\gamma\theta^\ell + \sum_{i\in S^1} s_i^\ell \le \gamma\beta\ell + 3n\sigma. \tag{4}$$

By Theorem 2.15, with probabilty at least $1 - 2e^{-\frac{\lambda^2\alpha}{280\ell}}$,

$$\gamma\theta^\ell + \sum_{i\in S^1} s_i^\ell \ge \frac{2}{\pi}(1-2\zeta)(1-2n\alpha)\ell - \lambda \tag{5}$$

for all $\lambda \in [0, 17.5\ell/\sqrt{\alpha}]$. Set $\lambda = \left(\frac{1}{2}-\beta\right)^2 \ell/2\pi$ and assme Equation (5) also holds.

Combining Equations (4) and (5) gives

$$\frac{2}{\pi}(1-2\zeta)(1-2n\alpha)\ell - \frac{\left(\frac{1}{2}-\beta\right)^2}{2\pi}\ell \le \gamma\beta\ell + 3n\sigma. \tag{6}$$

We claim this is a contradiction, which then holds with high probability, thus proving the theorem.

Rearranging Equation (6) gives

$$\frac{2}{\pi}(1-2\zeta)(1-2n\alpha) \le \frac{\left(\frac{1}{2}-\beta\right)^2}{2\pi} + \gamma\beta + \frac{3n\sigma}{\ell}. \tag{7}$$

Our setting of parameters gives

$$2n\alpha \le \frac{\left(\frac{1}{2}-\beta\right)}{2} \quad \text{and} \quad \frac{3n\sigma}{\ell} \le \frac{\left(\frac{1}{2}-\beta\right)^2}{2\pi}.$$

Substituting these into Equation (7) gives

$$\frac{2}{\pi}(1-2\zeta)\left(1 - \frac{1}{2}\left(\frac{1}{2}-\beta\right)\right) \le \frac{\left(\frac{1}{2}-\beta\right)^2}{\pi} + \gamma\beta. \tag{8}$$

Now we use $1 - 2\zeta = \frac{1}{2}\left(\frac{1}{2}-\beta\right)$ and $\gamma = \frac{2}{\pi}\frac{1-2\zeta}{\zeta} = \frac{\left(\frac{1}{2}-\beta\right)}{\pi\zeta}$ to derive a contradiction from Equation (8):

$$\begin{aligned}
\frac{\left(\frac{1}{2}-\beta\right)}{\pi}\left(1 - \frac{1}{2}\left(\frac{1}{2}-\beta\right)\right) &\le \frac{\left(\frac{1}{2}-\beta\right)^2}{\pi} + \frac{\left(\frac{1}{2}-\beta\right)}{\pi\zeta}\beta, \\
1 - \frac{1}{2}\left(\frac{1}{2}-\beta\right) &\le \left(\frac{1}{2}-\beta\right) + \frac{\beta}{\zeta}, \\
\zeta\left(1 - \frac{3}{2}\left(\frac{1}{2}-\beta\right)\right) &\le \beta.
\end{aligned}$$

Since $\zeta = \frac{1}{2} - \frac{1}{4}\left(\frac{1}{2} - \beta\right)$, we have

$$\zeta\left(1 - \frac{3}{2}\left(\frac{1}{2} - \beta\right)\right) = \frac{1}{2}\left(1 - \frac{1}{2}\left(\frac{1}{2} - \beta\right)\right)\left(1 - \frac{3}{2}\left(\frac{1}{2} - \beta\right)\right) > \frac{1}{2}\left(1 - 2\left(\frac{1}{2} - \beta\right)\right).$$

And

$$\beta = \frac{1}{2}\left(1 - 2\left(\frac{1}{2} - \beta\right)\right).$$

This gives a contradiction. The total failure probability is bounded by

$$e^{-n\sigma^2/4\ell} + e^{-\sqrt{\alpha}n\sigma/4} + 2e^{-\lambda^2\alpha/280\ell} \leq \left(\frac{\delta}{32}\right)^{16n} + \left(\frac{\delta}{32}\right)^{4n} + 2\left(\frac{\delta}{32}\right)^{\frac{1}{2}\left(\frac{1}{2}-\beta\right)n} \leq \delta^{\frac{1}{2}\left(\frac{1}{2}-\beta\right)n},$$

assuming $\left(\frac{1}{2} - \beta\right)n \geq 1$. □

3 Hardness of False Discovery

3.1 The Statistical Query Model

Given a distribution $\mathcal{D}$ over $\{0,1\}^d$, we would like to answer *statistical queries* about $\mathcal{D}$. A statistical query on $\{0,1\}^d$ is specified by a function $q : \{0,1\}^d \to [-1,1]$ and (abusing notation) is defined to be

$$q(\mathcal{D}) = \mathop{\mathbb{E}}_{x \leftarrow_R \mathcal{D}}[q(x)].$$

Our goal is to design an *oracle* $\mathcal{O}$ that answers statistical queries on $\mathcal{D}$ using only iid samples $x_1, \ldots, x_n \leftarrow_R \mathcal{D}$. Our focus is the case where the queries are chosen adaptively and adversarially.

Specifically, $\mathcal{O}$ is a stateful algorithm that holds a collection of samples $x_1, \ldots, x_n \in \{0,1\}^d$, takes a statistical query q as input, and returns a real-valued answer $a \in [-1,1]$. We require that when $x_1, \ldots, x_n$ are iid samples from $\mathcal{D}$, the answer a is close to $q(\mathcal{D})$, and moreover that this condition holds for every query in an adaptively chosen sequence $q^1, \ldots, q^\ell$. Formally, we define the following game between an $\mathcal{O}$ and a stateful adversary $\mathcal{A}$.

$\mathcal{A}$ chooses a distribution $\mathcal{D}$ over $\{0,1\}^d$.
Sample $x_1, \ldots, x_n \leftarrow_R \mathcal{D}$, let $x = (x_1, \ldots, x_n)$.
For $j = 1, \ldots, \ell$
  $\mathcal{A}$ outputs a query q^j.
  $\mathcal{O}(x, q^j)$ outputs a^j.
  (As $\mathcal{A}$ and $\mathcal{O}$ are stateful, q^j and a^j may depend on the history $q^1, a^1, \ldots, q^{j-1}, a^{j-1}$.)

Figure 3: $\mathsf{Acc}_{n,d,\ell}[\mathcal{O}, \mathcal{A}]$

Definition 3.1 (Accuracy)**.** An oracle $\mathcal{O}$ is (α, β, γ)*-accurate for* ℓ *adaptively chosen queries given* n *samples in* $\{0,1\}^d$ if for every adversary $\mathcal{A}$,

$$\mathop{\mathbb{P}}_{\mathsf{Acc}_{n,d,\ell}[\mathcal{O},\mathcal{A}]}\left[\text{For } (1-\beta)\ell \text{ choices of } j,\ \left|\mathcal{O}(x, q^j) - q^j(\mathcal{D})\right| \leq \alpha\right] \geq 1 - \gamma.$$

As a shorthand, we will say that $\mathcal{O}$ is (α, β)*-accurate for* ℓ *queries* if for every $n, d \in \mathbb{N}$, $\mathcal{O}$ is $(\alpha, \beta, o_n(1))$-accurate for ℓ queries given n samples in $\{0,1\}^d$. Here, ℓ may depend on n and d and $o_n(1)$ is a function of n that tends to 0.

We are interested in oracles that are both accurate and computationally efficient. We say that an oracle $\mathcal{O}$ is *computationally efficient* if when given samples $x_1, \ldots, x_n \in \{0,1\}^d$ and a query $q : \{0,1\}^d \to [-1,1]$ it runs in time $\mathrm{poly}(n, d, |q|)$. Here q will be represented as a circuit that evaluates $q(x)$ and $|q|$ denotes the size of this circuit.

3.2 Encryption Schemes

Our attack relies on the existence of a semantically secure private-key encryption scheme. An encryption scheme is a triple of efficient algorithms (Gen, Enc, Dec) with the following syntax:

- Gen is a randomized algorithm that takes as input a security parameter λ and outputs a λ-bit secret key. Formally, $sk \leftarrow_R Gen(1^\lambda)$.
- Enc is a randomized algorithm that takes as input a secret key and a message $m \in \{-1, 0, 1\}$ and outputs a ciphertext ct. Formally, $ct \leftarrow_R Enc(sk, m)$.
- Dec is a deterministic algorithm that takes as input a secret key and a ciphertext ct and outputs a decrypted message m'. If the ciphertext ct was an encryption of m under the key sk, then $m' = m$. Formally, if $ct \leftarrow_R Enc(sk, m)$, then $Dec(sk, ct) = m$ with probability 1.

Roughly, security of the encryption scheme asserts that no polynomial time adversary who does not know the secret key can distinguish encryptions of $m = 0$ from encryptions of $m = 1$, even if the adversary has access to an oracle that returns the encryption of an arbitrary message under the unknown key. For convenience, we will require that this security property holds simultaneously for an arbitrary polynomial number of secret keys. The existence of an encryption scheme with this property follows immediately from the existence an ordinary semantically secure encryption scheme. We start with the stronger definition only to simplify our proofs. A secure encryption scheme exists under the minimal cryptographic assumption that one-way functions exist. The formal definition of security is not needed until Section A.

3.3 Description of the Attack

The adversary is specified in Figure 4. In Figure 4, (Gen, Enc, Dec) is an encryption scheme and $\mathcal{F}$ is an n-collusion resilient interactive fingerprinting code for N users of length $\ell = \ell(N)$ robust to a β fraction of errors with false accusation probability $\delta = 1/8$. Observe that $\mathsf{Attack}_{n,d}$ is only well defined for pairs $n, d \in N$ for which $1 + \lceil \log_2(2000n) \rceil \le d$, so that there exists a suitable choice of $\lambda \in \mathbb{N}$. Through this section we will assume that $n = n(d)$ is a polynomial in d and that d is a sufficiently large unspecified constant, which ensures that $\mathsf{Attack}_{n,d}$ is well defined.

3.4 Analysis of the Attack

We will start by establishing that the number of falsely accused users is small. That is, letting ψ^ℓ denote the number of users in T^ℓ who are not in the set S, we have $\psi^\ell \le N/8$ with high probability. This condition will follow from the security of the interactive fingerprinting code $\mathcal{F}$. However, security alone is not enough to guarantee that the number of falsely accused users is small, because security of $\mathcal{F}$ applies to adversaries that only have access to c_i^j for users $i \in S \setminus T^j$, whereas the queries to the oracle depend on c_i^j for users $i \notin S \setminus T^j$. To remedy this

The distribution $\mathcal{D}$:
  Given parameters d, n, let $N = 2000n$, let $\lambda = d - \lceil \log_2(2000n) \rceil$.
  For $i \in [N]$, let $sk_i \leftarrow_R Gen(1^\lambda)$ and let $y_i = (i, sk_i) \in \{0,1\}^d$.
  Let $\mathcal{D}$ be the uniform distribution over $\{y_1, \ldots, y_N\} \subseteq \{0,1\}^d$.

Choose samples $x_1, \ldots, x_n \leftarrow_R \mathcal{D}$, let $x = (x_1, \ldots, x_n)$.
Let $S \subseteq [N]$ be the set of unique indices i such that (i, sk_i) appears in x.

Attack:
  Let $T^1 = \emptyset$.
  For $j = 1, \ldots, \ell = \ell(N)$:
    Let $c^j \in \{\pm 1\}^N$ be the column given by $\mathcal{F}$.
    For $i = 1, \ldots, N$, let $ct_i^j = Enc(sk_i, c_i^j)$.
    Define the query $q^j(i', sk')$ to be $Dec(sk', ct_{i'}^j)$ if $i' \notin T^j$ and 0 otherwise.
    Let $a^j = \mathcal{O}(x; q^j)$ and round a^j to $\{\pm 1\}$ to obtain $\overline{a}^j$.
    Give $\overline{a}^j$ to $\mathcal{F}$ and let $I^j \subseteq [N]$ be the set of accused users and $T^j = T^{j-1} \cup I^j$.

Figure 4: $\mathsf{Attack}_{n,d}[\mathcal{O}]$

problem we rely on the fact entries c_i^j for i outside of $S \setminus T^j$ are encrypted under keys sk_i that are not known to the oracle. Thus, a computationally efficient oracle "does not know" those rows. We can formalize this argument by comparing Attack to an IdealAttack (Figure 5) where these entries are replaced with zeros, and argue that the adversary cannot distinguish between these two attacks without breaking the security of the encryption scheme.

Claim 3.2. *For every oracle $\mathcal{O}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}\left[\psi^\ell > N/8\right] \leq \mathrm{negl}(n)$$

Proof. This follow straightforwardly from a reduction to the security of the fingerprinting code. Notice that since the query q^j does not depend on any entry c_i^j for $i \notin S \setminus T^{j-1}$. Thus, an adversary for the fingerprinting code who has access to $c^j_{S \setminus T^{j-1}}$ can simulate the view of the oracle. Since we have for any adversary $\mathcal{P}$

$$\mathop{\mathbb{P}}_{\mathsf{IFPC}_{N,n,\ell}[\mathcal{P},\mathcal{F}]}\left[\psi^\ell > N/8\right] \leq \mathrm{negl}(n),$$

we also have

$$\mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}\left[\psi^\ell > N/8\right] \leq \mathrm{negl}(n),$$

as desired. □

Now we can argue that an efficient oracle cannot distinguish between the real attack and the ideal attack. Thus the conclusion that $\psi^\ell \leq N/8$ with high probability must also hold in the real game.

The distribution $\mathcal{D}$:
Given parameters d, n, let $N = 2000n$, and $\lambda = d - \lceil \log_2(2000n) \rceil$.
For $i \in [N]$, let $sk_i \leftarrow_R Gen(1^\lambda)$ and let $y_i = (i, sk_i) \in \{0,1\}^d$.
Let $\mathcal{D}$ be the uniform distribution over $\{y_1, \ldots, y_N\} \subseteq \{0,1\}^d$.

Choose samples $x_1, \ldots, x_n \leftarrow_R \mathcal{D}$, let $x = (x_1, \ldots, x_n)$.
Let $S \subseteq [N]$ be the set of unique indices i such that (i, sk_i) appears in x.

Recovery phase:
Let $T^1 = \emptyset$.
For $j = 1, \ldots, \ell = \ell(N)$:
Let $c^j \in \{\pm 1\}^N$ be the column given by $\mathcal{F}$.
For $i \in S$, let $ct_i^j = Enc(sk_i, c_i^j)$, for $i \in [N] \setminus S$, let $ct_i^j = Enc(sk_i, 0)$.
Define the query $q^j(i', sk')$ to be $Dec(sk', ct_{i'}^j)$ if $i' \notin T^j$ and 0 otherwise.
Let $a^j = \mathcal{O}(x; q^j)$ and round a^j to $\{\pm 1\}$ to obtain $\overline{a}^j$.
Give $\overline{a}^j$ to $\mathcal{F}$ and let $I^j \subseteq [N]$ be the set of accused users and $T^j = T^{j-1} \cup I^j$.

Figure 5: $\mathsf{IdealAttack}_{n,d}[\mathcal{O}]$

Claim 3.3. *Let Z_1 be the event $\left\{\psi^\ell > N/8\right\}$. Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then if $\mathcal{O}$ is computationally efficient, for every $d \in \mathbb{N}$*

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}[Z_1] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{O}]}[Z_1] \right| \leq \mathrm{negl}(n)$$

The proof is straightforward from the definition of security, and is deferred to Section A. Combining Claims 3.2 and 3.3 we easily obtain the following.

Claim 3.4. *For every computionally efficient oracle $\mathcal{O}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{O}]}\left[\psi^\ell > N/8\right] \leq \mathrm{negl}(n)$$

Claim 3.4 will be useful because it will allow us to establish that an accurate oracle must give answers that are consistent with the fingerprinting code. That is, using θ^ℓ to denote the number of inconsistent answers $\overline{a}^1, \ldots, \overline{a}^\ell$, we will have $\theta^\ell \ll \ell/2$ with high probability.

Claim 3.5. *If $\mathcal{O}$ is $(1/3, \beta)$-accurate for $\ell = \ell(2000n)$ adaptively chosen queries then for every polynomial $n = n(d)$ and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{O}]}\left[\theta^\ell \leq \beta\ell\right] \geq 1 - o_n(1)$$

Proof. In the attack, the oracle's input consists of n samples from $\mathcal{D}$, and the total number of queries issued is ℓ. Therefore, by the assumption that $\mathcal{O}$ is $(1/3, \beta)$-accurate for ℓ queries, we have

$$\mathbb{P}\left[\begin{array}{c} \text{For } (1-\beta)\ell \text{ choices of } j \in [\ell], \\ \left| \mathcal{O}(x, q^j) - \mathop{\mathbb{E}}_{(i,sk_i) \leftarrow_R \mathcal{D}}\left[q^j(i, sk_i)\right] \right| \leq 1/3 \end{array}\right] \geq 1 - o_n(1). \tag{9}$$

Observe that, by construction, for every $j \in [\ell]$,

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{(i,sk_i)\leftarrow_R \mathcal{D}} \left[q^j(i, sk_i) \right] - \mathop{\mathbb{E}}_{i\in[N]} \left[c_i^j \right] \right| \\
&= \left| \left(\frac{1}{N} \sum_{i\in[N]\setminus T^{j-1}} Dec(sk_i, ct_i^j) + \frac{1}{N} \sum_{i\in T^{j-1}} 0 \right) - \mathop{\mathbb{E}}_{i\in[n]} \left[c_i^j \right] \right| \\
&= \left| \left(\frac{1}{N} \sum_{i\in[N]\setminus T^{j-1}} c_i^j \right) - \mathop{\mathbb{E}}_{i\in[n]} \left[c_i^j \right] \right| \\
&\leq \frac{2\left|T^{j-1}\right|}{N} \\
&\leq \frac{2(\psi^{j-1}+n)}{N}
\end{aligned} \tag{10}$$

where the second equality is because by construction $ct_i^j \leftarrow_R Enc(sk_i, c_i^j)$ and the inequality is because we have $c_i^j \in \{\pm 1\}$.

By Claim 3.4, and the fact that $\psi^{j-1} \leq \psi^\ell$, we have

$$\mathbb{P}\left[\psi^{j-1} > N/8 + n\right] \leq \mathbb{P}\left[\psi^\ell > N/8 + n\right] \leq \mathrm{negl}(n).$$

Noting that $N/8 + n \leq N/6$ and combining with (10), we have

$$\mathbb{P}\left[\forall\ j \in [\ell],\ \left| \mathop{\mathbb{E}}_{(i,sk_i)\leftarrow_R \mathcal{D}} \left[q^j(i, sk_i) \right] - \mathop{\mathbb{E}}_{i\in[n]} \left[c_i^j \right] \right| \leq 1/3 \right] \geq 1 - \mathrm{negl}(n)$$

Applying the triangle inequality to (9) and (10), we obtain

$$\mathbb{P}\left[\begin{array}{c} \text{For } (1-\beta)\ell \text{ choices of } j \in [\ell], \\ \left| \mathcal{O}(x, q^j) - \mathop{\mathbb{E}}_{i\in[N]} \left[c_i^j \right] \right| \leq 2/3 \end{array} \right] \geq 1 - o_n(1). \tag{11}$$

Fix a $j \in [\ell]$ such that a^j is accurate for query q^j. If $c_i^j = 1$ for every $i \in S \setminus T^{j-1}$, then by (10), $a^j = \mathcal{O}(x, q^j) \geq 1/3$, so the rounded answer $\overline{a}^j = 1$. Similarly if $c_i^j = -1$ for every $i \in S \setminus T^{j-1}$, $\overline{a}^j = -1$. Therefore there must exist $i \in S \setminus T^{j-1}$ so that $\overline{a}^j = c_i^j$. Thus there are $(1-\beta)\ell$ choices of $j \in [\ell]$ for which this condition holds, so the number of errors θ^ℓ is at most $\beta\ell$. This completes the proof of the claim. □

As before, we can argue that the real attack and the ideal attack are computationally indistinguishable, and thus the oracle must also give consistent answers in the ideal attack.

Claim 3.6. *Let Z_2 be the event $\left\{\theta^\ell \leq \beta\ell\right\}$. Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then if $\mathcal{O}$ is computationally efficient, for every $d \in \mathbb{N}$*

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}[Z_2] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{O}]}[Z_2] \right| \leq \mathrm{negl}(n)$$

The proof is straightforward from the definition of security, and is deferred to Section A. Combining Claims 3.5 and 3.6 we easily obtain the following.

Claim 3.7. *If $\mathcal{O}$ computationally efficient and $(1/3,\beta)$-accurate for $\ell = \ell(2000n)$ adaptively chosen queries then for every polynomial $n = n(d)$ and every sufficiently large $d \in \mathbb{N}$,*

$$\mathbb{P}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}\left[\theta^{\ell} \leq \beta\ell\right] \geq 1 - o_n(1)$$

However, the conclusion of 3.7 can easily be seen to lead to a contradiction, because the security of the fingerprinting code assures that no attacker who only has access to $c^j_{S\setminus T^{j-1}}$ in each round $j = 1,\ldots,\ell$ can give answers that are consistent for $(1-\beta)\ell$ of the columns c^j. Thus, we have

Claim 3.8. *For every oracle $\mathcal{O}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$,*

$$\mathbb{P}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}\left[\theta^{\ell} \leq \beta\ell\right] \leq \mathrm{negl}(n)$$

Putting it together, we obtain the following theorem.

Theorem 3.9. *There is a function $\ell(2000n,\beta) = O(n^2/\left(\frac{1}{2}-\beta\right)^4)$ such that there is no computationally efficient oracle $\mathcal{O}$ that is $(1/3,\beta)$-accurate for $\ell(2000n,\beta)$ adaptively chosen queries given n samples in $\{0,1\}^d$.*

Proof. Assume for the sake of contradiction that there were such an oracle. Then by Claim 3.7 we would have

$$\mathbb{P}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}\left[\theta^{\ell} \leq \beta\ell\right] \geq 1 - o_n(1).$$

But, by Claim 3.8 we have

$$\mathbb{P}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}\left[\theta^{\ell} \leq \beta\ell\right] \leq \mathrm{negl}(n),$$

which is a contradiction. □

3.5 An Information-Theoretic Lower Bound

As in [HU14], we observe that the techniques underlying our computational hardness result can also be used to prove an information-theoretic lower bound when the dimension of the data is large. At a high level, the argument uses the fact that the encryption scheme we rely on only needs to satisfy relatively weak security properties, specifically security for at most $O(n^2)$ messages. This security property can actually be achieved against computationally unbounded adversaries provided that the length of the secret keys is $O(n^2)$. As a result, our lower bound can be made to hold against computationally unbounded oracles, but since the secret keys have length $O(n^2)$, we will require $d = O(n^2)$. We refer the reader to [HU14] for a slightly more detailed discussion, and simply state the following result.

Theorem 3.10. *There is a function $\ell(2000n,\beta) = O(n^2/\left(\frac{1}{2}-\beta\right)^4)$ such that there is no oracle $\mathcal{O}$ (even one that is computationally unbounded) that is $(1/3,\beta)$-accurate for $\ell(2000n,\beta)$ adaptively chosen queries given n samples in $\{0,1\}^d$ when $d \geq \ell(2000n,\beta)$.*

4 Hardness of Avoiding Blatant Non Privacy

In this section we show how our arguments also imply that computationally efficient oracles that guarantee accuracy for adaptively chosen statistical queries must be blatantly non-private.

4.1 Blatant Non Privacy and Sample Accuracy

Before we can define blatant non-privacy, we need to define a notion of accuracy that is more appropriate for the application to privacy. In contrast to Definition 3.1 where accuracy is defined with respect to the distribution, here we define accurate with respect to the sample itself. With this change in mind, we model blatant non-privacy via the following game.

$\mathcal{A}_{\text{priv}}$ chooses a set $y = \{y_1, \ldots, y_{2n}\} \subseteq \{0,1\}^d$
Sample a random subsample $x \subseteq_R y$ of size n
For $j = 1, \ldots, \ell$
  $\mathcal{A}_{\text{priv}}$ outputs a query q^j
  $\mathcal{O}(x, q^j)$ outputs a^j
  (As $\mathcal{A}_{\text{priv}}$ and $\mathcal{O}$ are stateful, q^j and a^j may depend on $q^1, a^1, \ldots, q^{j-1}, a^{j-1}$.)
$\mathcal{A}_{\text{priv}}$ outputs a set $x' \subseteq y$

Figure 6: $\mathsf{NonPrivacy}_{n,d}[\mathcal{O}, \mathcal{A}_{\text{priv}}]$

Definition 4.1. An oracle $\mathcal{O}$ is (α, β, γ)-*sample-accurate for* ℓ *adaptively chosen queries given* n *samples in* $\{0,1\}^d$ if for every adversary $\mathcal{A}_{\text{priv}}$,

$$\mathop{\mathbb{P}}_{\mathsf{NonPrivacy}_{n,d,\ell}[\mathcal{O},\mathcal{A}_{\text{priv}}]}\left[\text{For } (1-\beta)\ell \text{ choices of } j \in [\ell], \left|\mathcal{O}(x, q^j) - q^j(x)\right| \leq \alpha\right] \geq 1 - \gamma$$

where $q(x) = \frac{1}{n}\sum_{i\in[n]} q(x_i)$ is the average over the *sample*.

As a shorthand, we will say that $\mathcal{O}$ is (α, β)-*sample-accurate for* ℓ *queries* if for every $n, d \in \mathbb{N}$, $\mathcal{O}$ is $(\alpha, \beta, o_n(1))$-accurate for ℓ queries given n samples in $\{0,1\}^d$. Here, ℓ may depend on n and d and $o_n(1)$ is a function of n that tends to 0.

Definition 4.2. *Giving* (α, β)-*accurate answers to* ℓ *adaptively chosen queries is blatantly non-private for efficient oracles* if there exists an adversary $\mathcal{A}_{\text{priv}}$ such that for every oracle $\mathcal{O}$ that is computationally efficient and (α, β)-sample-accurate for ℓ adaptively chosen queries,

$$\mathop{\mathbb{P}}_{\mathsf{NonPrivacy}_{n,d,\ell}[\mathcal{O},\mathcal{A}_{\text{priv}}]}\left[|x \triangle x'| > n/100\right] \leq o_n(1)$$

If the conclusion holds even for computationally inefficient oracles then we replace "for efficient oracles" with "for unbounded oracles" in the definition.

4.2 Lower Bounds

In this section we show the following theorem

Theorem 4.3. *Giving accurate answers to* $O(n^2)$ *adaptively chosen queries is blatantly non-private for computationally efficient oracles.*

The attack is defined in Figure 7. Therein $\mathcal{F}$ is a n-collusion-resilient interactive fingerprinting code of length ℓ for $N = 2n$ users robust to a β fraction of errors with false accusation probability $\delta = 1/2000$.

The set y:
  Given parameters d, n, let $\lambda = d - \lceil \log_2(2n) \rceil$.
  For $i \in [2n]$, let $sk_i \leftarrow_R Gen(1^\lambda)$ and let $y_i = (i, sk_i)$.

Attack:
  Let $T^1 = \emptyset$.
  For $j = 1, \ldots, \ell = \ell(2n)$:
    Let $c^j \in \{\pm 1\}^{2n}$ be the column given by $\mathcal{F}$.
    For $i = 1, \ldots, 2n$, let $ct_i^j = Enc(sk_i, c_i^j)$.
    Define the query $q^j(i', sk')$ to be $Dec(sk', ct_{i'}^j)$ if $i' \notin T^j$ and 0 otherwise.
    Let $a^j = \mathcal{O}(x; q^j)$ and round $(n/(n - |T^{j-1}|))a^j$ to $\{\pm 1\}$ to obtain $\overline{a}^j$.
    Give $\overline{a}^j$ to $\mathcal{F}$ and let $I^j \subseteq [N]$ be the set of accused users and $T^j = T^{j-1} \cup I^j$.
    If $|T^j| > 499n/500$, let $L = j$, halt, and output $x' = \{y_i : i \in T^L\}$.
  Let $L = \ell$, and output $x' = \{y_i : i \in T^L\}$.

Figure 7: $\mathsf{PrivacyAttack}_{n,d}[\mathcal{O}]$

We will start by establishing that the number of falsely accused users is small. That is, letting ψ^j denote the number of users in T^j who are not in the set x, we have $\psi^L \leq n/1000$ with high probability. As in Section 3, this condition will follow from the security of the interactive fingerprinting code $\mathcal{F}$ combined with the security of the encryption scheme, via the introduction of an "ideal attack" (Figure 8).

The set y:
  Given parameters d, n, let $\lambda = d - \lceil \log_2(2n) \rceil$.
  For $i \in [2n]$, let $sk_i \leftarrow_R Gen(1^\lambda)$ and let $y_i = (i, sk_i)$.

Attack:
  Let $T^1 = \emptyset$.
  For $j = 1, \ldots, \ell = \ell(2n)$:
    Let $c^j \in \{\pm 1\}^{2n}$ be the column given by $\mathcal{F}$.
    For $i = 1, \ldots, 2n$, let $ct_i^j = Enc(sk_i, c_i^j)$.
    For $i \in S$, let $ct_i^j = Enc(sk_i, c_i^j)$, for $i \in [N] \setminus x$, let $ct_i^j = Enc(sk_i, 0)$.
    Let $a^j = \mathcal{O}(x; q^j)$ and round $(n/(n - |T^{j-1}|))a^j$ to $\{\pm 1\}$ to obtain $\overline{a}^j$.
    Give $\overline{a}^j$ to $\mathcal{F}$ and let $I^j \subseteq [N]$ be the set of accused users and $T^j = T^{j-1} \cup I^j$.
    If $|T^j| > 499n/500$, let $L = j$, halt, and output $x' = \{y_i : i \in T^L\}$.
  Let $L = \ell$, and output $x' = \{y_i : i \in T^L\}$.

Figure 8: $\mathsf{IdealPrivacyAttack}_{n,d}[\mathcal{O}]$

Claim 4.4. *For every oracle $\mathcal{O}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}\left[\psi^L > n/1000\right] \leq \mathrm{negl}(n)$$

Proof. This follow straightforwardly from a reduction to the security of the fingerprinting code. Notice that since the query q^j does not depend on any entry c_i^j for $i \notin x \setminus T^{j-1}$. Thus, an adversary for the fingerprinting code who has access to $c^j_{x\setminus T^{j-1}}$ can simulate the view of the oracle. Since we have for any adversary $\mathcal{P}$

$$\mathop{\mathbb{P}}_{\mathsf{IFPC}_{N,n,\ell}[\mathcal{P},\mathcal{F}]}\left[\psi^\ell > N/2000\right] \leq \mathrm{negl}(n),$$

we also have

$$\mathop{\mathbb{P}}_{\mathsf{IdealPrivacyAttack}_{n,d}[\mathcal{O}]}\left[\psi^L > n/1000\right] \leq \mathrm{negl}(n),$$

where we have used the fact that for every $j' \leq j$, $\psi^{j'} \leq \psi^j$. This completes the proof. □

Now we can argue that an efficient oracle cannot distinguish between the real attack and the ideal attack. Thus the conclusion that $\psi^L \leq n/1000$ with high probability must also hold in the real game.

Claim 4.5. *Let Z_1 be the event $\left\{\psi^L > n/1000\right\}$ Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then if $\mathcal{O}$ is computationally efficient, for every $d \in \mathbb{N}$*

$$\left|\mathop{\mathbb{P}}_{\mathsf{IdealPrivacyAttack}_{n,d}[\mathcal{O}]}[Z_1] - \mathop{\mathbb{P}}_{\mathsf{PrivacyAttack}_{n,d}[\mathcal{O}]}[Z_1]\right| \leq \mathrm{negl}(n)$$

The proof is straightforward from the definition of security, and is deferred to Section A. Combining Claims 4.4 and 4.5 we easily obtain the following.

Claim 4.6. *For every computionally efficient oracle $\mathcal{O}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{PrivacyAttack}_{n,d}[\mathcal{O}]}\left[\psi^L > n/1000\right] \leq \mathrm{negl}(n)$$

By Claim 4.6 we have $|x' \setminus x| \leq n/1000$. Now, in order to show $|x' \triangle x| \leq n/100$, it suffices to show that $|x \setminus x'| \leq n/200$. In order to do so we begin with the following claim, which establishes that if the oracle $\mathcal{O}$ is sufficiently accurate, and $|x \setminus T^{j-1}| \leq n/200$, then the oracle returns a consistent answer to the query q^j. Recalling that we use θ^j to denote the number of rounded answers $\overline{a}^k$ for $1 \leq k \leq j$ that are inconsistent with c^j, we can state the following claim.

Claim 4.7. *If $\mathcal{O}$ is $(1/1000, 0)$-sample-accurate for $\ell = \ell(2n)$ adaptively chosen queries then for every polynomial $n = n(d)$, every sufficiently large $d \in \mathbb{N}$, and every $j \in [L]$,*

$$\mathop{\mathbb{P}}_{\mathsf{PrivacyAttack}_{n,d}[\mathcal{O}]}\left[\theta^L = 0\right] \geq 1 - o_n(1)$$

Proof. Observe that, by construction, for every $j \in [\ell]$,

$$\begin{aligned}\mathop{\mathbb{E}}_{i\in x}\left[q^j(x_i)\right] &= \frac{1}{n}\left(\sum_{i\in(x\setminus T^{j-1})} c_i^j + \sum_{i\in(x\cap T^{j-1})} 0\right)\\ &= \mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right]\cdot\left(\frac{|x\setminus T^{j-1}|}{n}\right)\end{aligned}$$

After renormalizing by $(n/n-|T^{j-1}|)$ we have

$$\begin{aligned}&\left(\frac{n}{n-|T^{j-1}|}\right)\cdot\mathop{\mathbb{E}}_{i\in x}\left[q^j(x_i)\right]\\ &= \mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right]\cdot\left(\frac{n}{n-|T^{j-1}|}\right)\cdot\left(\frac{|x\setminus T^{j-1}|}{n}\right)\\ &= \mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right]\cdot\left(\frac{n-T^{j-1}+\psi^{j-1}}{n-|T^{j-1}|}\right)\\ &= \mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right]\cdot\left(1+\frac{\psi^{j-1}}{n-|T^{j-1}|}\right)\end{aligned}$$

Since $0 \le \psi^{j-1} \le n/10000$ (by Claim 4.6), and since the algorithm terminates unless $|T^{j-1}| \le 499n/500$, we obtain

$$\begin{aligned}&\mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right] \le \left(\frac{n}{n-|T^{j-1}|}\right)\cdot\mathop{\mathbb{E}}_{i\in x}\left[q^j(x_i)\right] \le \frac{11}{10}\cdot\mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right]\\ \implies &\left|\left(\frac{n}{n-|T^{j-1}|}\right)\cdot\mathop{\mathbb{E}}_{i\in x}\left[q^j(x_i)\right]-\mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right]\right| \le \frac{1}{10} \qquad (12)\end{aligned}$$

By the assumption that $\mathcal{O}$ is $(1/1000,0)$-sample-accurate, we have that, with probability $1-o_n(1)$, for every $j\in[L]$,

$$\left|a^j - \mathop{\mathbb{E}}_{i\in x}\left[q^j(x_i)\right]\right| \le 1/1000. \qquad (13)$$

Now, combining (12) and (13), we have

$$\begin{aligned}&\left|\left(\frac{n}{n-|T^{j-1}|}\right)\cdot a^j - \mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right]\right|\\ \le &\left|\left(\left(\frac{n}{n-|T^{j-1}|}\right)\cdot\mathop{\mathbb{E}}_{i\in x}\left[q^j(x_i)\right]-\mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right]\right)+\left(\frac{n}{n-|T^{j-1}|}\right)\cdot\frac{1}{1000}\right| \le \frac{1}{2}+\frac{1}{10}\le\frac{2}{3} \qquad (14)\end{aligned}$$

Finally, observe that if $c_i^j = 1$ for every $i\in[2n]$, then we have

$$\mathop{\mathbb{E}}_{i\in(x\setminus T^{j-1})}\left[c_i^j\right] = 1,$$

and by (14) we have $(n/(n-|T^j|))a^j \ge 1-2/3 = 1/3$. Thus, the rounded answer $\overline{a}^j = 1$. Similarly, if $c_i^j = -1$ for every $i\in[2n]$, then we have $\overline{a}^j = -1$. This completes the proof of the claim. □

As before, we can argue that the real attack and the ideal attack are computationally indistinguishable, and thus the oracle must also give consistent answers in the ideal attack.

Claim 4.8. *Let Z_2 be the event $\left\{\theta^L = 0\right\}$ Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then if $\mathcal{O}$ is computationally efficient, for every $d \in \mathbb{N}$*

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealPrivacyAttack}_{n,d}[\mathcal{O}]}[Z_2] - \mathop{\mathbb{P}}_{\mathsf{PrivacyAttack}_{n,d}[\mathcal{O}]}[Z_2] \right| \leq \mathrm{negl}(n)$$

The proof is straightforward from the definition of security, and is deferred to Section A. Combining Claims 4.7 and 4.8 we easily obtain the following.

Claim 4.9. *If $\mathcal{O}$ computationally efficient and $(1/1000, 0)$-accurate for $\ell = \ell(2n)$ adaptively chosen queries then for every polynomial $n = n(d)$ and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{IdealPrivacyAttack}_{n,d}[\mathcal{O}]}\left[\theta^L = 0\right] \geq 1 - o_n(1)$$

We can use Claim 4.9 to derive a contradiction. To do so we use the fact that the security of the fingerprinting code assures that no attacker who only has access to $c^j_{x \setminus T^{j-1}}$ in each round $j = 1, \ldots, \ell$ can give answers that are consistent for all ℓ of the columns c^j. Thus, we have

Claim 4.10. *For every oracle $\mathcal{O}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$, if $L = \ell$*

$$\mathop{\mathbb{P}}_{\mathsf{IdealPrivacyAttack}_{n,d}[\mathcal{O}]}\left[\theta^\ell = 0\right] \leq \mathrm{negl}(n)$$

Putting it together, we obtain the following theorem.

Theorem 4.11. *There is a function $\ell(2n) = O(n^2)$ such that there is no computationally efficient oracle $\mathcal{O}$ that is $(1/1000, 0)$-accurate for $\ell(2n)$ adaptively chosen queries given n samples in $\{0,1\}^d$.*

Proof. Assume for the sake of contradiction that there were such an oracle. Now consider two cases. First consider the case that $L < \ell$, which means the algorithm has terminated early due to the condition $|T^L| \geq 499n/500$ being reached. In this case we have $|x'| = |T^L| \geq 499n/500$. However, by Claim 4.4, we have that $|x' \setminus x| \leq n/10000$. Therefore we have $|x \triangle x'| \leq (499/500 - 1/5000)n \leq n/100$, as desired.

Now consider the case where $L = \ell$, meaning the algorithm does not terminate early. In this case, by Claim 4.9 we have

$$\mathop{\mathbb{P}}_{\mathsf{IdealPrivacyAttack}_{n,d}[\mathcal{O}]}\left[\theta^\ell = 0\right] \geq 1 - o_n(1)$$

but by Claim 4.10 we have

$$\mathop{\mathbb{P}}_{\mathsf{IdealPrivacyAttack}_{n,d}[\mathcal{O}]}\left[\theta^\ell = 0\right] \leq \mathrm{negl}(n),$$

which is a contradiction. This completes the proof of the theorem. □

4.3 An Information-Theoretic Lower Bound

As we did in Section 3.5, we can prove an information-theoretic analogue of our hardness result for avoiding blatant non-privacy.

Theorem 4.12. *There is a function $\ell(2n) = O(n^2)$ such that there is no oracle $\mathcal{O}$ (even a computationally unbounded one) that is $(1/1000, 0)$-accurate for $\ell(2n)$ adaptively chosen queries given n samples in $\{0,1\}^d$ where $d \geq \ell(2n)$.*

The proof is essentially identical to what is sketched in Section 3.5.

Acknowledgements

We thank Moritz Hardt and Salil Vadhan for insightful discussions during the early stages of this work.

References

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[Bon36] Carlo Emilio Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze.*, 8, 1936.

[BS98] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

[BUV14] Mark Bun, Jonathan Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, pages 1–10. ACM, May 31 – June 3 2014.

[CFN94] Benny Chor, Amos Fiat, and Moni Naor. Tracing traitors. In *CRYPTO*, pages 257–270. Springer, August 21-25 1994.

[DFH+14] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Guilt-free data exploration (tentative title). *Manuscript*, 2014.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, March 4-7 2006.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210. ACM, June 9-12 2003.

[DNR+09] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil P. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *STOC*, pages 381–390. ACM, May 31 - June 2 2009.

[Dun61] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64, 1961.

[Ete85] N. Etemadi. On some classical results in probability theory. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 47(2):pp. 215–221, 1985.

[FT01] Amos Fiat and Tamir Tassa. Dynamic traitor tracing. *J. Cryptology*, 14(3):211–223, 2001.

[HU14] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*. IEEE, October 19-21 2014.

[Kea93] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. In *STOC*, pages 392–401. ACM, May 16-18 1993.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[Tar08] Gábor Tardos. Optimal probabilistic fingerprint codes. *J. ACM*, 55(2), 2008.

[Ull13] Jonathan Ullman. Answering $n^{2+o(1)}$ counting queries with differential privacy is hard. In *STOC*, pages 361–370. ACM, June 1-4 2013.

[Ull14] Jonathan Ullman. Private multiplicative weights beyond linear queries. *CoRR*, abs/1407.1571, 2014.

A Security Reductions from Sections 3 and 4

In Section 3 we made several claims comparing the probability of events in Attack to the probability of events in IdealAttack. Each of these claims follow from the assumed security of the encryption scheme. In this section we restate and prove these claims. Since the claims are all of a similar nature, the proof will be somewhat modular. The claims in Section 4 relating PrivacyAttack to IdealPrivacyAttack can be proven in an essentially identical fashion, and we omit these proofs for brevity.

Before we begin recall the formal definition of security of an encryption scheme. Security is defined via a pair of oracles $\mathcal{E}_0$ and $\mathcal{E}_1$. $\mathcal{E}_1(sk_1,\ldots,sk_N,\cdot)$ takes as input the index of a key $i \in [N]$ and a message m and returns $Enc(sk_i, m)$, whereas $\mathcal{E}_0(sk_1,\ldots,sk_N,\cdot)$ takes the same input but returns $Enc(sk_i, 0)$. The security of the encryption scheme asserts that for randomly chosen secret keys, no computationally efficient adversary can tell whether or not it is interacting with $\mathcal{E}_0$ or $\mathcal{E}_1$.

Definition A.1. An encryption scheme (Gen, Enc, Dec) is *secure* if for every polynomial $N = N(\lambda)$, and every poly(λ)-time adversary $\mathcal{B}$, if $sk_1,\ldots,sk_N \leftarrow_R Gen(1^\lambda)$

$$\left|\mathbb{P}\left[\mathcal{B}^{\mathcal{E}_0(sk_1,\ldots,sk_N,\cdot)} = 1\right] - \mathbb{P}\left[\mathcal{B}^{\mathcal{E}_1(sk_1,\ldots,sk_N,\cdot)} = 1\right]\right| = \mathrm{negl}(\lambda)$$

We now restate the relevant claims from Section 3.

Claim A.2 (Claim 3.3 Restated)**.** *Let Z_1 be the event $\left\{\psi^\ell > N/8\right\}$. Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then if $\mathcal{O}$ is computationally efficient, for every $d \in \mathbb{N}$*

$$\left|\mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}[Z_1] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{O}]}[Z_1]\right| \leq \mathrm{negl}(n)$$

Claim A.3 (Claim 3.6 Restated). *Let Z_2 be the event $\left\{\theta^\ell \le \beta\ell\right\}$. Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then if $\mathcal{O}$ is computationally efficient, for every $d \in \mathbb{N}$*

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}[Z_2] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{O}]}[Z_2] \right| \le \mathrm{negl}(n)$$

To prove both of these claims, for $c \in \{1,2\}$ we construct an adversary $\mathcal{B}_c$ that will attempt to use $\mathcal{O}$ to break the security of the encryption. We construct $\mathcal{B}_c$ in such a way that its advantage in breaking the security of encryption is precisely the difference in the probability of the event Z_c between Attack and IdealAttack, which implies that the difference in probabilities is negligible. The simulator is given in Figure 9

Simulate constructing and sampling from $\mathcal{D}$:
  Given parameters d, n, let $N = 2000n$, let $\lambda = d - \lceil \log_2(2000n) \rceil$.
  Sample users $u_1, \ldots, u_n \leftarrow_R [N]$, let S be the set of distinct users in the sample.
  Choose new keys $sk_i \leftarrow_R Gen(1^\lambda)$ for $i \in S$.
  For $i \in S$, let $x_i = (u_i, sk_{u_i})$, let $x = (x_1, \ldots, x_n)$.

Simulate the attack:
  Let $T^1 = \emptyset$.
  For $j = 1, \ldots, \ell = \ell(N)$:
    Let c^j be the column given by $\mathcal{F}$.
    For $i = 1, \ldots, N$:
      If $i \in S$, let $ct_i^j = Enc(sk_i, c_i^j)$, otherwise as $\mathcal{E}$ for an encryption of c_i^j under key $\overline{sk}_i$, that is $ct_i^j = \mathcal{E}_b(\overline{sk}_1, \ldots, \overline{sk}_N, i, c_i^j)$.
    Define the query $q^j(i', sk')$ to be $Dec(sk', ct_{i'}^j)$ if $i' \notin T^j$ and 0 otherwise.
    Let $a^j = \mathcal{O}(x; q^j)$ and round a^j to $\{\pm 1\}$ to obtain $\overline{a}^j$.
    Give $\overline{a}^j$ to $\mathcal{F}$ and let $I^j \subseteq [N]$ be the set of accused users and $T^j = T^{j-1} \cup I^j$.

Output 1 if and only if the event Z_c occurs

Figure 9: $\mathcal{B}_{c,n,d}^{\mathcal{E}_b(\overline{sk}_1, \ldots, \overline{sk}_N, \cdot)}$

Proof of Claims A.2, A.3. First, observe that for $c \in \{1,2\}$, $\mathcal{B}_c$ is computationally efficient as long as $\mathcal{F}$ and $\mathcal{O}$ are both computationally efficient. It is not hard to see that our construction $\mathcal{F}$ is efficient and efficiency of $\mathcal{O}$ is an assumption of the claim. Also notice $\mathcal{B}$ can determine whether Z_c has occurred efficiently.

Now we observe that when the oracle is $\mathcal{E}_1$ (the oracle that takes as input i and m and returns $Enc(\overline{sk}_i, m)$), and $\overline{sk}_1, \ldots, \overline{sk}_N$ are chosen randomly from $Gen(1^\lambda)$, then the view of the oracle is identical to $\mathsf{Attack}_{n,d}[\mathcal{O}]$. Specifically, the oracle holds a random sample of pairs (i, sk_i) and is shown queries that are encryptions either under keys it knows or random unknown keys. Moreover, the messages being encrypted are chosen from the same distribution. On the other hand, when the oracle is $\mathcal{E}_0$ (the oracle that takes as input i and ct and returns $Enc(\overline{sk}_i, 0)$), then

the view of the oracle is identical to $\mathsf{Attack}_{n,d}[\mathcal{O}]$. Thus we have that for $c \in \{1,2\}$,

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{O}]}[Z_c] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{O}]}[Z_c] \right|$$

$$= \left| \mathop{\mathbb{P}}_{\overline{sk}_1,\ldots,\overline{sk}_N \leftarrow_{\mathrm{R}} Gen(1^\lambda)} \left[\mathcal{B}_{c,n,d}^{\mathcal{E}_0(\overline{sk}_1,\ldots,\overline{sk}_N,\cdot)} = 1 \right] - \mathop{\mathbb{P}}_{\overline{sk}_1,\ldots,\overline{sk}_N \leftarrow_{\mathrm{R}} Gen(1^\lambda)} \left[\mathcal{B}_{c,n,d}^{\mathcal{E}_1(\overline{sk}_1,\ldots,\overline{sk}_N,\cdot)} = 1 \right] \right| = \mathrm{negl}(\lambda) = \mathrm{negl}(d)$$

The last equality holds because we have chosen $N = 2000n(d) = \mathrm{poly}(d)$, and therefore we have $\lambda = d - \lceil \log N \rceil = d - O(\log d)$. This completes the proof of both claims. □