Incompressible Encodings

Tal Moran[1] and Daniel Wichs[2⋆]

[1] IDC Herzliya. talm@idc.ac.il
[2] Northeastern University and NTT Research Inc. wichs@ccs.neu.edu

Abstract. An *incompressible encoding* can probabilistically encode some data m into a codeword c, which is not much larger. Anyone can decode the codeword c to recover the original data m. However, the codeword c cannot be efficiently compressed, even if the original data m is given to the decompression procedure on the side. In other words, c is an efficiently decodable representation of m, yet is computationally incompressible even given m. An incompressible encoding is *composable* if many encodings cannot be simultaneously compressed.

The recent work of Damgård, Ganesh and Orlandi (CRYPTO '19) defined a variant of incompressible encodings as a building block for "proofs of replicated storage". They constructed incompressible encodings in an ideal permutation model, but it was left open if they can be constructed under standard assumptions, or even in the more basic random-oracle model. In this work, we undertake the comprehensive study of incompressible encodings as a primitive of independent interest and give new constructions, negative results and applications:

- We construct incompressible encodings in the *common random string (CRS)* model under either Decisional Composite Residuosity (DCR) or Learning with Errors (LWE). However, the construction has several drawbacks: (1) it is not composable, (2) it only achieves selective security, and (3) the CRS is as long as the data m.
- We leverage the above construction to also get a scheme in the random-oracle model, under the same assumptions, that avoids all of the above drawbacks. Furthermore, it is significantly more efficient than the prior ideal-model construction.
- We give black-box separations, showing that incompressible encodings in the plain model cannot be proven secure under any standard hardness assumption, and incompressible encodings in the CRS model must inherently suffer from all of the drawbacks above.
- We give a new application to "big-key cryptography in the bounded-retrieval model", where secret keys are made intentionally huge to make them hard to exfiltrate. Using incompressible encodings, we can get all the security benefits of a big key without wasting storage space, by having the key to encode useful data.

⋆ Research supported by NSF grants CNS-1314722, CNS-1413964, CNS-1750795 and the Alfred P. Sloan Research Fellowship.

1 Introduction

The entire contents of Wikipedia can be downloaded as a compressed (gzip) file that is several gigabytes in size. Can it be compressed further? This is an interesting question, and there is much work on optimizing compression for various types of data. But there is also an uninteresting answer – the Wikipedia contents can be easily compressed via the link "www.wikipedia.org", which allows anyone on the Internet to recover the data! We consider the compression problem in exactly the above scenario, where the data is readily available publicly. However, our goal is to make such data incompressible. That is, we want to come up with an *incompressible encoding* scheme that takes some such data (e.g., Wikipedia content) and probabilistically represents it in a way that is guaranteed to be incompressible, even when the decompression procedure is given the original data on the side for free (e.g., can access Wikipedia over the Internet). We now elaborate on what this primitive is and why it is useful.

Incompressible Encodings. An incompressible encoding scheme consists of a pair of efficient and key-less encoding and decoding procedures $(\mathsf{Enc}, \mathsf{Dec})$, such that $c \leftarrow \mathsf{Enc}(m)$ probabilistically encodes some data m into a codeword c that anybody can efficiency decode to recover $m = \mathsf{Dec}(c)$. The size of the codeword c should not be "much larger" than that of the underlying data m. Even though the codeword c is an efficiently decodable representation of the data m, and cannot contain much additional information since its size is not much larger, we want c to be incompressible even given m. In particular, we consider the following *incompressibility* security game:

1. An adversary chooses some arbitrary data m, which is encoded to get a codeword $c \leftarrow \mathsf{Enc}(m)$.
2. The codeword c is given to an adversarial compression algorithm that outputs some compressed value w.
3. The compressed value w along with underlying data m are given to an adversarial decompressor, who wins if it outputs the codeword c.

We require that no efficient adversary can win the above game with better than negligible probability, unless w is "almost as large" as the entire codeword.

Good vs. Trivial Encodings. The definition of incompressible encodings has two parameters: for data $m \in \{0,1\}^k$, we let $\alpha(k)$ denote the *codeword size*, and $\beta(k)$ denote the *incompressibility parameter*, which bounds size of the value w output by the compressor in the security game.

For a "good" incompressible encoding, we want the codeword size to be essentially the same as the size of the underlying data $\alpha(k) = (1 + o(1))k$ and we want the encoded data to be incompressible below the size of essentially the entire data/codeword $\beta(k) = (1 - o(1))k = (1 - o(1))\alpha(k)$.[3]

[3] Throughout the introduction, we will ignore additive polynomial terms in the security parameter, which means that the above bounds only hold when the data size k is sufficiently large.

On the other hand, there is a "trivial" incompressible encoding, which just appends some randomness r to the data and sets the codeword to $c = (m, r)$. Since r cannot be compressed, the encoding also cannot be compressed below the length of r. This ensures that $\alpha(k) = k + |r|$ and $\beta(k) = |r|$. Therefore a scheme with $\alpha(k) \geq k + \beta(k)$ is "trivial" and we say that an incompressible encoding is "non-trivial" if it beats this bound.

It is easy to see that non-trivial incompressible encodings cannot be constructed information theoretically. This is because, there are at most $2^{\alpha(k)-k}$ possible codewords per message on average, and therefore also certainly for the worst-case message m. A pair of inefficient compression/decompression procedures can enumerate the list of all such codewords (e.g., in lexiographic order) and compress/decompress any codeword in the list just by writing down its index using $\beta(k) = \alpha(k) - k$ bits. Therefore, we will need to rely on computational assumptions to construct non-trivial incompressible encodings.

Local Decoding. We will also want our incompressible encodings to be *locally decodable*, so that any bit of the message can be recovered by only reading a few bits of the codeword. While our constructions have this property, most of the challenges are already present even without this requirement.

Composability. We say that an incompressible encoding scheme, with some incompressibility parameter β, is *composable* if any n independently generated encodings of various messages cannot be compressed below $\sum_{i\in[n]} \beta(k_i)$ bits, where k_i denotes the length of the i'th message. Moreover, the probability of the adversarial compression algorithm outputting less than $\sum_{i\in\mathcal{I}} \beta(k_i)$ bits and the decompression procedure reconstructing all of the codewords $c_i \ : \ i \in \mathcal{I}$ for some set $\mathcal{I} \subseteq [n]$ should be negligible. As we will discuss later, we do not know whether all incompressible encodings are inherently composable (we conjecture otherwise) and therefore we define it as a separate security property of interest.

Keyed vs Key-less Schemes. We mention that any semantically secure encryption scheme would give an incompressible encoding where the decoding procedure needs a secret key. This is because the encryption of some arbitrary data m is indistinguishable from an encryption of random data, which is inherently incompressible even given m. The main difficulty of our problem is that we require the encoding and decoding procedures to be public and key-less; anybody should be able to decode the codeword to recover the data.

1.1 Prior and Concurrent Work

Proofs of Replicated Storage. The work of Damgård, Ganesh and Orlandi [17] studied a primitive called "Proofs of Replicated Storage", which considers a setting where we want to store some data m with a cloud storage provider, who promises to store n replicated copies of the data at different locations ("replicas") to increase fault tolerance. We want to be able to check that the provider is indeed storing n copies of the data. To solve this problem they defined a novel building block called a "replica encoding", which corresponds to our notion of an incompressible encodings that is also *bounded self-composable*; i.e., it is composable when the same message is encoded n times for some a-priori bound n. Using such encodings, they construct "proofs of replicated storage" by encoding the data n separate times and having each replica store a different codeword. This is combined with "Proof of Retreivability" [22,35,51], which are used to periodically audit each replica and check that it is storing its codeword. The cloud provider cannot meaningfully save on storage while maintaining the ability to pass the audit with non-negligible probability. "Proofs of Replicated Storage" also have applications to the *Filecoin* cryptocurrency [37,38], where replicated storage of user data is a resource for mining coins; see [17] for details. We note that some prior notions of "proofs of replicated storage" and "incompressible/replica encodings" were also previously considered by [9,13,24] in settings where the adversary is computationally more constrained than the encoding/decoding procedures and does not have the ability to run these procedures; here we focus on the setting where the adversary can run in arbitrary polynomial time.

Construction of Incompressible Encodings of [17]. In the above context, the work of Damgård, Ganesh and Orlandi [17] (building on an idea from an earlier work of [52]) proposed a construction of incompressible (replica) encodings based on a family of trapdoor permutations (TDPs) denoted by f_{pk} and an ideal invertible public random permutation P, P^{-1}. To encode a message m, the encoder samples a random TDP public key and trapdoor $(\mathsf{pk}, \mathsf{td})$ and outputs a codeword $c = (\mathsf{pk}, (f_{\mathsf{pk}}^{-1} \circ P)^r(m))$ by applying the function $g(x) = f_{\mathsf{pk}}^{-1}(P(x))$ iteratively for r rounds starting with the message m, where r is some parameter. The codeword is efficiently decodable by computing f_{pk} in the forward direction and making calls to P^{-1}. While the above requires using a TDP whose domain is as large as the entire file, it is also possible to apply the TDP on smaller blocks of the file separately. Unfortunately, the construction has several big deficiencies:

- We discovered that the security proof in the published version of [17] is fundamentally flawed. In fact, we identified some heuristic counter-examples to suggest that their scheme is unlikely to be secure in general with the number of rounds r claimed. However, we conjectured (and it has since been confirmed by a concurrent work, see below) that the scheme can be proved secure when the number of rounds r is made very large: $r = \Omega(kn)$ where k is the file-size (in bits), n is the number of compositions and λ is the security parameter. Since each round takes $\Omega(k)$ time, this means that the scheme has complexity $\Omega(k^2 n)$ and, in particular, runs in time quadratic in the file size, which we envision to be large. Moreover, the scheme needs to perform essentially that many public key operations (RSA exponentiations), which makes the scheme highly impractical. Furthermore, the scheme only achieves bounded-composability where the number of compositions n needs to be known ahead of time and affects the complexity of the scheme.
- The construction works in the "ideal invertible random permutation" model. Furthermore, the domain/range of the ideal permutation has to match that

of the TDP. For example, if we use the RSA TDP, then the domain/range of P needs to be $\mathbb{Z}_N^*$ where N is the RSA modulus. It remained as an open problem to give a construction in the standard model, or even in the random oracle model, or even using an ideal permutations where the domain/range is just $\{0,1\}^s$ for some parameter s. (While it is possible to construct indifferentiable [41] ideal permutations from a random oracle [16], they will not have the required structured domain/range. Furthermore, the indifferentiability framework is insufficient to guarantee security for multi-stage games [49], which includes the security game for incompressible encodings.)

Concurrent Work of [26]. A concurrent and independent work of Garg, Lu and Waters [26] independently discovered the flaw in the proof of [17]. They managed to patch the result by providing a completely new (and highly involved) proof of security for their scheme, when the number of rounds r is made sufficiently large $r = \Omega(kn)$ as discussed above. They also managed to remove the reliance on an "ideal invertible random permutation" with a structured domain and showed how to instantiate the scheme using the random oracle model alone.

1.2 Our Results

Our work initiates the study of incompressible encodings as a primitive of independent interest. We give new positive and negative results as well as a new application of this primitive as follows.

Constructions in the CRS and RO Model. We give a new construction of incompressible encodings in the *common random string (CRS)* model under either the *Decisional Composite Residuosity (DCR)* or the *Learning with Errors (LWE)* assumptions. Our construction relies on certain types of *lossy trapdoor functions* (LTFs) [47] that are also permutations/surjective, and our constructions of these may be of independent interest. The encodings have good parameters with codeword size $\alpha(k) = (1 + o(1))k$ and incompressibility parameter $\beta(k) = (1 - o(1))k = (1 - o(1))\alpha(k)$. Furhtermore, the encoding/decoding runtime only scales linearly in the data-size. The scheme is also locally decodable. However, it suffers from three drawbacks:

1. It is *not composable* if all the encodings use the same CRS (but is composable if each encoding gets a fresh CRS).
2. It only achieves *selective security*, where the message being encoded cannot be chosen adaptively depending on the CRS.
3. It has a *long CRS*, linear in the length of the data m.

We also leverage our construction in the CRS model to get a construction in the *random-oracle (RO) model* under the same assumptions that avoids all of the above drawbacks. Namely, it is fully composable without any a priori bound on the number of compositions n, and achieves adaptive security, where the adversary can choose the message after making random-oracle queries.

Our schemes represent a significant improvement over the construction and analysis of [17, 26] since:

– We get the first constructions in the CRS model without relying on an ideal object (ideal permutation or random oracle), albeit with the above-mentioned drawbacks.
– Our schemes in both the CRS and RO models are significantly more efficient and our encoding/decoding run time is $O(k)$ rather than $O(k^2)$, where k is the data size. Since we generally envision encoding large data where k is several gigabytes or terabytes, the difference between linear and quadratic run-time is highly significant. (We omit factors in the security parameter in the above bounds).
– Our RO scheme is fully composable and we do not need to bound the number of compositions n ahead of time nor does the efficiency of the scheme degrade with n.
– We can plausibly achieve post-quantum security via our LWE-based construction whereas [17, 26] seemed to inherently require trapdoor permutations for which we have no good candidates with post-quantum security.
– Our proof of security is in many ways significantly simpler than that of [26].

Black-Box Separations. We give black-box separations, showing that non-trivial incompressible encodings in the plain model cannot be proven secure via a black-box reduction from any standard hardness assumption, including strong assumptions such as the existence of indistinguishability obfuscation. Moreover, we show that similar black-box separations apply to good incompressible encodings in the CRS model, unless they suffer from all 3 of the above drawbacks: they cannot be fully composable with a single CRS, they cannot achieve adaptive security, and the CRS needs to be essentially as long as the data m. This shows that our results are in some sense optimal.

Application to Big-Key Crypto. We give a new application of incompressible encodings to "big-key cryptography in the bounded-retrieval model" [2, 3, 8, 12, 19, 23, ...]. Big-key cryptosystems are designed with intentionally huge secret keys in order to make them hard to exfiltrate. They guarantee security even if the adversary can get large amounts of arbitrary "leakage" on the secret key.

Using incompressible encodings, we can get all the security benefits of a big key without wasting storage space, by allowing the key to encode useful data. We do not need to assume that the data has any entropy from the point of view of the adversary; for example the user's secret key could encode the contents of Wikipedia, or the user's movie collection, or other data that a user may want to store offline but is also publicly available on the Internet and may be fully known to the adversary.

In particular, we show how to construct public-key encryption in this model, where the secret key is an incompressible encoding of the user's arbitrary data. Security is maintained even if the adversary can exfiltrate arbitrary information about the secret key, as long as the size of such leakage is bounded by some $(1 - o(1))$ fraction of the secret key size. The public key size, ciphertext size and the encryption/decryption run time are all small and only poly-logarithmic in the data size. In particular, each decryption operation only accesses some small subset of the secret key bits.

1.3 Our Techniques

We now delve into each of the results above and the relevant techniques in turn.

Incompressible Encodings in the CRS model. We construct incompressible encodings in the CRS model. In this model, the honest encoding/decoding algorithms as well as the adversarial compression/decompression algorithms have access to a public uniformly random string. Our construction relies on certain types of *lossy tradpor functions* (LTFs) [47]. An LTF consists of a family of function f_{pk} indexed by a public key pk. The public key can be sampled in one of two indistinguishable modes: in *injective mode* the function f_{pk} is injective and we can sample pk along with a trapdoor td that allows us to efficiently invert it, and in *lossy mode* the function f_{pk} has a very small image, meaning that $f_{\mathsf{pk}}(x)$ reveals very little information about the input x.

As a starting point, to illustrate the main ideas, let's assume that we have an LTF family $f_{\mathsf{pk}} : \{0,1\}^d \to \{0,1\}^d$ where both the domain and range are equal to $\{0,1\}^d$ for some polynomial d. This means that f_{pk} is a permutation over $\{0,1\}^d$ in injective mode. Let's also assume that the LTF is highly lossy, meaning that the output $f_{\mathsf{pk}}(x)$ in lossy mode only reveals $o(d)$ bits of information about x. To encode some data $m \in \{0,1\}^k$ we think of $m = (m_1, \dots, m_{k'})$ as consisting of $k' = k/d$ blocks of length d each. We also rely on a common random string $\mathsf{crs} = (u_1, \dots, u_{k'})$ consisting of k' random blocks of length d each. The encoding procedure $\mathsf{Enc}_{\mathsf{crs}}(m)$ samples a random pk in injective mode, together with a trapdoor td and sets the codeword to be $c = (\mathsf{pk}, f_{\mathsf{pk}}^{-1}(m_1 \oplus u_1), \dots, f_{\mathsf{pk}}^{-1}(m_{k'} \oplus u_{k'})$. The decoding procedure $\mathsf{Dec}_{\mathsf{crs}}(c)$ recovers m by applying f_{pk} in the forward direction and xor'ing out the u_i components. Moreover, it's easy to see that individual locations of the data can be decoded locally. The codeword is of size $k + |\mathsf{pk}| = (1 + o(1))k$.

We prove incompressible security of the above scheme in the selective setting, where the choice of the message m is worst-case but cannot depend on the CRS. We first observe that the joint distribution of $\mathsf{crs} = (u_1, \dots, u_{k'}), c = (\mathsf{pk}, x_1, \dots, x_{k'})$ sampled as above with $u_i \leftarrow \{0,1\}^d$ and $x_i = f_{\mathsf{pk}}^{-1}(m_i \oplus u_i)$ is identical to the distribution where we choose $x_i \leftarrow \{0,1\}^d$ uniformly at random and set $u_i = f_{\mathsf{pk}}(x_i) \oplus m_i$. The latter distribution can be sampled without a trapdoor. Therefore, we can indistinguishably switch pk to lossy mode. Now the codeword c has $(1 - o(1))k$ bits of true entropy even conditioned on crs and m, since each of the values $u_i = f_{\mathsf{pk}}(x_i) \oplus m_i$ reveals only $o(d)$ bits of information about each x_i. Therefore, we can argue that in this case the codeword c is (even information-theoretically) incompressible below $(1 - o(1))k$ bits, even given crs and m. But since this case is computationally indistinguishable from the real distribution of crs, c, the same must hold computationally there as well.

To give some more intuition on the above idea, note that in reality the codeword c has very little actual entropy given crs, m. This is inherent since we want to c to be almost the same size as m and it has to decode to m so there is no space left to inject actual entropy. But in the security proof, we indistinguishably move the information about the message m into the crs and allow the codeword to have a high amount of real entropy even given crs, m while preserving the condition that it decodes to m. Therefore, we can argue incompressibility.

Surjective Lossy Functions. There are many constructions of lossy trapdoor functions (LTFs) in the literature (e.g.,) [6,25,36,47,55]. However, the vast majority of them are not *surjective* and hence are unusable for our scheme where we need to compute inverses $f_{\mathsf{pk}}^{-1}(m_1 \oplus u_1)$ on arbitrary values that we cannot force to be in the image. Fortunately, the work of [25] gives a construction of a surjective LTF (permutation) based on Paillier's Decisional Composite Residuosity (DCR) assumption [44] using the ideas behind the Damgård-Jurik cryptosystem [18]. Furthermore, this LTF can be made highly lossy to ensure that the output only reveals an $o(1)$ fraction of the information in the input. There is still some subtlety in that the domain and range of the LTF are not bit-strings $\{0,1\}^d$ but rather the group $\mathbb{Z}^*_{N^{s+1}}$ for some RSA modulus N and some parameter s. It turns out that we can nevertheless use it in our construction with minor modifications, while maintaining a uniformly random CRS. This is because we can use truly random bits in the CRS to obliviously sample random elements in the group $\mathbb{Z}^*_{N^{s+1}}$.

Since surjective LTFs are also trapdoor permutations (TDPs), all good candidates rely on assuming the hardness of factoring and it is a long-standing open problem to get constructions from other assumptions, such as DDH or LWE. However, we notice that we don't need our functions to be *injective*. Instead we define a relaxation of surjective LTFs that we call *Surjective Lossy Functions (SLFs)*. SLFs have two modes: a surjective (but not necessarily injective) mode and a lossy mode. In surjective mode, the image of the function f_{pk} is the entire range while in lossy mode the image is much smaller. Furthermore, in surjective mode we require a trapdoor that allows us to sample a random preimage of any value in the range. If the function is surjective but not injective, the input-length must be larger than the output-length, and we will require that it is at most $(1 + o(1))$ times larger to ensure that our encodings have small α. We show that such SLFs suffice in our construction of incompressible encodings. We then proceed to construct such SLFs under the learning with errors (LWE) assumption [48]. Our starting point is the surjective trapdoor function of [1,27,42] defined as $f_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{Z}_q^{n \times m}$ and $\mathbf{x} \in \mathbb{Z}_q^m$ has "small" entries. Unfortunately, in the best known instantiations [42], the input length (in bits) is at least twice as large as the output length, while we only want a $(1 + o(1))$ increase. We show a new technique to get around this. The high level idea is somewhat similar to a recent work of [14], and we rely on "approximate trapdoors" where, given $\mathbf{y}$, we can sample $\mathbf{x}$ such that $\mathbf{A}\mathbf{x}$ is close but not identical to $\mathbf{y}$. We modify the function to $f_{\mathbf{A}}(\mathbf{x}) = \lceil \mathbf{A}\mathbf{x} \rfloor_p$ by applying some rounding to the output to get rid of this difference. This allows us to optimize the ratio of input-size to output-size and get an improvements over exact trapdoors. To ensure a $(1 + o(1))$ overhead, our instantiation of this idea is somewhat different and arguably simpler than that of [14]. Furthermore, we also show that this function has a lossy mode where the output only reveals an $o(1)$ fraction of the information in the input, using the techniques of [4, 29]. Overall, we believe that this construction may be of independent interest.

Composability and HILL vs Yao Entropies. We cannot prove our construction of incompressible encodings in the CRS model to be composable if the same CRS is used to generate all the encodings. However, we show that it is composable if each encoding is given a fresh CRS. But first, let us discuss why composability is generally difficult.

One may be tempted to conjecture that *all* incompressible encodings are inherently composable. For example, it seems intuitive that if the adversary needs to store $\beta(k)$ bits to compress one codeword, she would need $2\beta(k)$ bits to compress two codewords of two length k messages (potentially the same message). Surprisingly, this intuition does not naturally translate into a proof — how would one design a reduction which leverages a compression algorithm that takes two codewords and compressed them below 2β bits to compress a single codeword below β bits? The situation is highly reminiscent of the "leakage amplification" problem in leakage-resilient cryptography: if some cryptosystem is secure even given β bits of leakage on its secret key, does that mean that two copies of the cryptosystem cannot be simultaneously broken even given 2β bits of leakage on the two keys? Surprisingly this is not the case and it was possible to construct clever counter-examples showing that the above does not generically hold in some cases [20, 34, 40]. We conjecture that counter-examples may also exist for incompressible encodings, at least under strong enough assumptions, and leave it as an open problem to come up with one.

Given that it is unknown (and perhaps unlikely) that all incompressible encodings are composable, we leverage a special property of our construction to show composability. In particular, the definition of incompressible encodings essentially says that c has high "Yao incompressibility entropy" [7, 33, 54] even given crs. However, for our construction, we showed that c even has a high HILL entropy [32] given crs, since the distribution of (crs, c) is computationally indistinguishable from one, where c has true (statistical) entropy given crs. Having HILL entropy is a stronger property than having incompressibility entropy and we leverage this to prove composition. In particular, while we do not know if the "Yao incompressibility entropy" of independent samples adds up, we do know that this is the case for HILL entropy: i.e., the HILL entropy of independent samples (crs_i, c_i) for $i \in [n]$ is the sum of the individual HILL entropies. This property implies composability for our encodings.

Construction in the Random Oracle Model. Our CRS model construction needed a long CRS, and a fresh CRS for each encoding if we wanted to argue composition. Furthermore, it only achieved selective security. We show how to resolve all of these drawbacks in the random oracle model. In this model all honest and adversarial algorithms have access to a truly random (fixed-length) function $\mathsf{RO} : \{0,1\}^\lambda \times \{0,1\}^\lambda \to \{0,1\}^\lambda$, where λ is the security parameter. The idea behind our construction is simple: the encoding procedure chooses fresh short randomness $r \leftarrow \{0,1\}^\lambda$ on each execution and then calls $\mathsf{RO}(r,1), \mathsf{RO}(r,2), \ldots$ to expand it into an arbitrarily long crs as needed; it then uses the incompressible encoding scheme in the CRS model with the above crs to encode the data and appends r to the codeword. This essentially corresponds to using a fresh CRS for each encoding. Note that, in the random oracle model, we allow the adversary to choose the message adaptively after making random oracle queries. However, we only need to rely on selective security in the CRS model since the adversary cannot predict r ahead of time and hence the crs used in each encoding is essentially freshly chosen after the adversary selects the message.

Black-Box Separations. We give black-box separations, showing that incompressible encodings in the plain model cannot be proven secure under any standard hardness assumption, and incompressible encodings in the CRS model must inherently suffer from all of the drawbacks that ours has. Our black-box separations have a similar flavor to those of [53] and we first explain their framework tailored to our setting.

They define the class of "single-stage game" assumptions, which are modeled via a game between a (potentially inefficient) challenger and a single stateful adversary; the assumption states that any polynomial-time adversary should have at most a negligible advantage in winning the game. This captures essentially all standard assumptions used in cryptography, from the hardness of factoring to indistinguishability obfuscation (iO).[4] We observe that the security definition of incompressible encodings is *not* a single-stage game since it involves two separate entities (the compressor and the decompressor) who cannot fully share state or communicate with each other — the compressor is limited in the number of bits it can pass to the decompressor. This makes it possible to separate the security of incompressible encodings from all single-stage game assumptions.

The separation works by constructing a "simulatable attacker". This is an *inefficient* (exponential size) attacker $\mathcal{A} = (\mathcal{A}.\mathsf{Compress}, \mathcal{A}.\mathsf{Expand})$ that breaks incompressible-encoding security. However we also design an efficient simulator $\mathcal{A}'$, such that one cannot (statistically) distinguish between black-box access to $\mathcal{A}$ versus $\mathcal{A}'$. Unlike the attackers $\mathcal{A}$, the simulator $\mathcal{A}'$ is a single fully stateful entity that can fully remember any inputs for invocations of $\mathcal{A}'.\mathsf{Compress}$ and use them to answer future invocations of $\mathcal{A}'.\mathsf{Expand}$. Therefore $\mathcal{A}'$ is not a legal attacker against the incompressible encoding. However, if some reduction can break a single-stage game assumption given black-box access to the legal (but inefficient) $\mathcal{A}$ then it would also be able to do so given access to the efficient (but illegal) $\mathcal{A}'$ which means that the assumption is false.

On a very high level, our adversary $\mathcal{A}$ uses "brute-force" to find the index i of the codeword in the lexicographic ordering of all codewords that decode to m and applies a random permutation on i to get the compressed value w. The decompressor inverts the permutation on w to recover i and uses that to

[4] This is a larger class than falsifiable assumptions [28, 43], where the challenger is also required to be efficient.

recover the codeword. The efficient simulator $\mathcal{A}'$ just chooses a fresh random w to compress each codeword but keeps a table of codewords it has seen with the corresponding w values it gave. To decompress given w, it just finds the corresponding codeword in the table.

In the above, we need to argue that the brute-force approach always works and that the number of codewords that decode to m is small so that the index i does not require too many bits to transmit. This holds in the plain model, when we choose the worst case message, or even in the CRS model when the message m can be chosen randomly but after the CRS if fixed. However, it fails in the selective-security setting in the CRS model – as it should, since we have a construction! This is because $\mathcal{A}$ may be given a CRS for which there are too many codewords that decode to the message m and hence it cannot write down the index i. If $\mathcal{A}$ failed in these cases but succeeded otherwise, we could use it to distinguish between such a CRS and a truly random one, which is something that cannot be efficiently simulated and indeed allows for a security reduction under standard assumptions.

Application to Big-Key Crypto. We give a new application of incompressible encodings to "big-key cryptography in the bounded-retrieval model" [2,3,8,12,19,23], where secret keys are intentionally huge to make them hard to exfiltrate. In particular, these works envision cryptosystems where the keys are many gigabytes or terabytes in size. Even if an adversary hacks into the system storing the secret key and manages to leak out large amounts of arbitrary data (but sufficiently less than the key size), we want the security of the cryptosystem to be maintained. For example, in the case of encryption, this should not enable the adversary to break semantic security of any ciphertexts sent in the future (unfortunately, the adversary can always perform decryption on past ciphertexts on the compromised device and leak out some plaintext bits so we cannot guarantee security of past ciphertexts). In such schemes, we want to maintain efficiency of honest users even as the key grows, and therefore the cryptosystem cannot even read the entire key to perform basic operations such as encryption and decryption. Prior work focused on constructing various primitives in this setting including symmetric-key encryption, public-key encryption and authenticated key agreement.

One big disadvantage of big-key cryptography is that it forces honest users to "waste" space by storing a huge random key of the cryptosystem. In this work, we propose to get rid of this waste by converting useful data that the user would store anyway into a cryptographic key. For example, the user can take their local movie/music collection (or an offline copy of Wikipedia etc.) and turn it into a cryptographic key for a "big-key" cryptosystem. We do not assume that the underlying data has any entropy from the point of view of the attacker; for example, the movie/music collection may be easily compressible into a very short list of titles that are available for download on the Internet. In general, we allow the attacker to choose the data in a worst-case fashion. This seems to make such data unusable as a cryptographic key, since it may be completely known to the adversary! However, this is where incompressible encodings come in. We ask the user to store an encoded version of the data and use the codeword as the key. The user can still access the data since it can be efficiently (locally) decoded. However, an adversary cannot easily exfiltrate the key by compressing it, even if the underlying data is completely known.

Turning the above high-level idea into workable big-key cryptosystems presents two challenges:

– We need to rely on big-key cryptosystems where the secret key can be an arbitrary string, rather than cryptosystems that choose a carefully structured secret key. This is especially a challenge for public-key encryption (PKE) schemes, where keys tend to have a lot of structure.
– We need to rely on a big-key cryptosystem that is secure with leakage, as long as the secret key is incompressible, even if it is not uniformly random. In particular, any attack against the leakage resilience of the encryption scheme should translate into a compression/decompression procedure on the secret key.

We call such cryptosystems *encoding-friendly*, and it is easy to see that they remain secure if we set their key to be an incompressible encoding of some arbitrary data.

We construct an *encoding-friendly* big-key PKE in the bounded-retrieval model, where the secret key can be arbitrarily large but the public key, ciphertexts, and encryption/decryption complexity are all small, only poly-logarithmic in the key size. Our work departs significantly from the prior construction of big-key PKE in the bounded-retrieval model of [2], which was not encoding friendly (the secret key in that scheme consisted of many "identity secret keys" in a special identity-based encryption scheme and hence required a large amount of structure). Instead, our construction relies on *laconic oblivious transfer (LOT)* [15], which can in turn be constructed under a wide range of assumptions such as CDH, LWE, and Factoring [11]. In an LOT scheme, one can take a long string $x \in \{0,1\}^k$ and hash it down into a short digest which acts as a public-key $\mathsf{pk} = h(x)$. One can then encrypt some arbitrary data μ under the public key pk with respect to a tuple $(i,b) \in [k] \times \{0,1\}$ such that the resulting ciphertext $\mathsf{ct} \leftarrow \mathsf{Enc}_{\mathsf{pk}}(\mu,(i,b))$ can be decrypted correctly using x if the i'th bit of x is $x[i] = b$. On the other hand, if $x[i] = 1 - b$ then the encryption is semantically secure even given x.

In our construction of an encoding-friendly big-key PKE, we can take an arbitrary (big) value x as the secret key and define the corresponding (short) public key as $\mathsf{pk} = h(x)$. To encrypt a message μ, we apply a secret sharing to derive random shares $\mu_1, \ldots, \mu_\lambda$ that sum up to μ, where λ is the security parameter. We then choose λ random indices $i_1, \ldots, i_\lambda$ and create 2λ LOT ciphertexts $\mathsf{ct}_{j,0} \leftarrow \mathsf{Enc}_{\mathsf{pk}}(\mu_j,(i_j,0))$ and $\mathsf{ct}_{j,1} \leftarrow \mathsf{Enc}_{\mathsf{pk}}(\mu_j,(i_j,1))$ where $\mathsf{ct}_{j,0}$ can be decrypted if $x[i_j] = 0$ and $\mathsf{ct}_{j,1}$ can be decrypted if $x[i_j] = 1$ respectively. We send all the ciphertexts along with the indices i_j. The decryption algorithms runs the LOT decryption on the ciphertexts $\mathsf{ct}_{j,x[i_j]}$ using the secret key x.

To prove security, we argue that we can extract (almost all of) x from any successful distinguisher given the leakage. We do this by choosing random ciphertexts and for each j testing if the distinguisher's advantage goes down noticeably when we replace the component $\mathsf{ct}_{j,0}$ by an encryption of random junk: if it does then we learn $x[i_j] = 0$ (since otherwise he could not notice this change) and otherwise we know $x[i_j] = 1$ (since if his advantage remains high, he must be decrypting $\mathsf{ct}_{j,1}$). By doing this for sufficiently many random ciphertexts we can recover a $(1 - o(1))$ fraction of the bits of x. This gives us a way to compresss x, by writing down the leakage together with a few additional bits that we can't recover from the distinguisher, and decompress it by running the distinguisher.

Organization. Due to lack of space, our black-box separation results, the application to big-key cryptography, and some proofs are deferred to the full version.

2 Preliminaries

Throughout, we let λ denote the security parameter. By default, all our statements hold in the non-uniform model of computation and we define PPT algorithms as circuits of size polynomial in their input and the security parameter. For $n \in \mathbb{Z}$ we let $[n]$ denote the set $[n] = \{1, \ldots, n\}$. When X is a distribution, or a random variable following this distribution, we let $x \leftarrow X$ denote the process of sampling x according to the distribution X. If $\mathcal{X}$ is a set, we let $x \leftarrow \mathcal{X}$ denote sampling x uniformly at random from $\mathcal{X}$.

Let X, Y be random variables, potentially parameterized by the security parameter. We define their *statistical difference* as $\mathsf{SD}(X, Y) = \frac{1}{2}\sum_u |\Pr[X = u] - \Pr[Y = u]|$. We write $X \overset{s}{\approx} Y$ if the statistical distance is negligible in the security parameter. We write $X \overset{c}{\approx} Y$ if they are computationally indistinguishable: for all PPT distinguishers D we have $|\Pr[D(X) = 1] - Pr[D(Y) = 1]| \leq \mathrm{negl}(\lambda)$.

The *min-entropy* of a random variable X is $H_\infty(X) = -\log(\max_x \Pr[X = x])$. Following Dodis et al. [21], we define the (average) conditional min-entropy of X given Y as: $H_\infty(X|Y) = -\log\left(\mathbb{E}_{y \leftarrow Y}\left[2^{-H_\infty(X|Y=y)}\right]\right)$. Note that $H_\infty(X|Y) = k$ iff the optimal strategy for guessing X given Y succeeds with probability 2^{-k}.

Lemma 1. *For any random variables X, Y where Y is supported over a set of size T we have $H_\infty(X|Y) \leq H_\infty(X) - \log T$.*

3 Defining Incompressible Encodings

We begin by giving a definition of incompressible encodings. Our first definition does not require composability and can be thought of as a simplified version of the replica encoding definition of [17].

Definition 1. *An (α, β)-incompressible encoding scheme consists of PPT algorithms* $(\mathsf{Enc}, \mathsf{Dec})$*. We require the following properties:*

Correctness: *There is some negligible μ such that for all $\lambda \in \mathbb{N}$ and all $m \in \{0,1\}^*$ we have* $\Pr[\mathsf{Dec}(\mathsf{Enc}(1^\lambda, m)) = m] = 1 - \mu(\lambda)$.

α-Expansion: *For all $\lambda, k \in \mathbb{N}$ and all $m \in \{0,1\}^k$ we have* $\Pr[|\mathsf{Enc}(1^\lambda, m)| \leq \alpha(\lambda, k)] = 1$.

β-Incompressibility: *Consider the following "compression experiment"* $\mathsf{CompExp}_{\mathcal{A}}(1^\lambda)$ *with an adversary* $\mathcal{A} = (\mathcal{A}.\mathsf{Select}, \mathcal{A}.\mathsf{Compress}, \mathcal{A}.\mathsf{Expand})$*:*
- $(m, \mathsf{aux}) \leftarrow \mathcal{A}.\mathsf{Select}(1^\lambda)$.
- $c \leftarrow \mathsf{Enc}(1^\lambda, m)$.
- $w \leftarrow \mathcal{A}.\mathsf{Compress}(\mathsf{aux}, c)$.
- $c' \leftarrow \mathcal{A}.\mathsf{Expand}(\mathsf{aux}, w)$.
- *Output 1 if* $c = c'$ *and* $|w| \leq \beta(\lambda, |m|)$.

We require that for all PPT $\mathcal{A}$ we have $\Pr[\mathsf{CompExp}_{\mathcal{A}}(1^\lambda) = 1] = \mathrm{negl}(\lambda)$.

We also refer to a "good" incompressible encoding, without specifying parameters (α, β) to refer to an (α, β)-incompressible encoding where $\alpha(\lambda, k) = k(1 + o(1)) + \mathrm{poly}(\lambda)$ and $\beta(\lambda, k) = k(1 - o(1)) - \mathrm{poly}(\lambda) = \alpha(\lambda, k)(1 - o(1)) - \mathrm{poly}(\lambda)$.

Composability. We also define a stronger notion of *composable* incompressible encodings. This can be thought of as a generalization of the replica encoding definition of [17], which only required "self-composition" when the same message was encoded many times.

Definition 2. *An (α, β)-incompressible encoding is* composable *if the following holds. Consider the following "composable compression experiment"* $\mathsf{CCExp}_{\mathcal{A}}(1^\lambda)$ *with an adversary* $\mathcal{A} = (\mathcal{A}.\mathsf{Select}, \mathcal{A}.\mathsf{Compress}, \mathcal{A}.\mathsf{Expand})$*:*

- $(\{m_i\}_{i=1}^n, \mathsf{aux}) \leftarrow \mathcal{A}.\mathsf{Select}(1^\lambda)$.
- *For* $i = 1 \ldots, n$*:* $c_i \leftarrow \mathsf{Enc}(1^\lambda, m_i)$.
- $w \leftarrow \mathcal{A}.\mathsf{Compress}(\mathsf{aux}, \{c_i\}_{i=1}^n)$.
- $\{c_i'\}_{i=1}^n \leftarrow \mathcal{A}.\mathsf{Expand}(\mathsf{aux}, w)$.
- *Let* $\mathcal{I} = \{i \ : \ c_i' = c_i\}$. *Output 1 if* $\mathcal{I} \neq \emptyset$ *and* $|w| \leq \sum_{i \in \mathcal{I}} \beta(\lambda, |m_i|)$.

We require that for all PPT $\mathcal{A}$ we have $\Pr[\mathsf{CCExp}_{\mathcal{A}}(1^\lambda) = 1] = \mathrm{negl}(\lambda)$.

CRS Model. We can generalize the above definitions to the *common random string* (CRS) model, where the encoding/decoding algorithms as well as the adversary $\mathcal{A}$ are given a random string $\mathsf{crs} \leftarrow \{0,1\}^{t(\lambda,k)}$ as an input. The length $t(\lambda, k)$ of the crs may depend on the message length k. In the CRS model, we distinguish between *selective security* and *adaptive security*. For selective security, the crs is not given to $\mathcal{A}.\mathsf{Select}$ but is given to $\mathcal{A}.\mathsf{Compress}, \mathcal{A}.\mathsf{Expand}$, meaning that the adversary's choice of the data m cannot depend on the crs. For adaptive security, $\mathcal{A}.\mathsf{Select}$ is also given the crs, and therefore the choice of m can depend on the crs. We say a scheme is selectively (resp. adaptively) β-incompressible in the CRS model.

Random Oracle Model. We can also generalize the above definitions to the *random-oracle model*, where the encoding/decoding algorithms as well as the adversarial algorithms $\mathcal{A}.\mathsf{Select}$, $\mathcal{A}.\mathsf{Compress}$, $\mathcal{A}.\mathsf{Expand}$ are given oracle access to a truly random function $\mathsf{RO} : \{0,1\}^\lambda \times \{0,1\}^\lambda \to \{0,1\}^\lambda$. Note that, in the random oracle model, we automatically require adaptive security, where the adversary's choice of the message m can adaptively depend on its random-oracle queries.

Locally Decodable. An incompressible encoding scheme $(\mathsf{Enc}, \mathsf{Dec})$ is locally decodable if there is some additional local decoding procedure that can recover any bit of the encoded message in time that is only poly-logarithmic in the message length k. In particular, we require that there is an algorithm $\mathsf{LocalDec}^c(1^\lambda, k, i)$ that gets RAM access to the input c and runs in time $\text{poly}(\lambda, \log k)$ such that the following holds. There is a negligible μ such that for any $\lambda, k \in \mathbb{N}$, any $m = (m_1, \ldots, m_k) \in \{0,1\}^k$ and any $i \in [k]$ we have $\Pr[\mathsf{LocalDec}^c(1^\lambda, k, i) = m_i \ : \ c \leftarrow \mathsf{Enc}(1^\lambda, m)] = 1 - \mu(\lambda)$.

Note on (Im)Perfect Correctness. While it will be convenient to allow imperfect correctness (with negligible failure probability) in our definition, we note that we can take any scheme with imperfect correctness and convert it into one with perfect correctness at a slight loss of security. The idea is to modify the encoding algorithm to also tests whether decoding succeeds: if so it outputs the codeword with a 0 appended to it and otherwise it outputs the message in the clear with a 1 appended to it. The decoding checks the appended bit and either applies the original decoding if it is a 0 or just outputs the rest of the codeword if the bit is a 1. If correctness of the original scheme holds with overwhelming probability than the above transformation cannot harm security.

Note on Message Selection. In the above, we allow an adversary $\mathcal{A}.\mathsf{Select}$ to choose the message m along with some auxiliary information aux (which can without loss of generality include the message m). For non-uniform adversaries, we can assume that $\mathcal{A}.\mathsf{Select}$ is deterministic (by hard-coding the worst-case choice of its random coins). For incompressibility in the plain model or with selective security in the CRS model, this means that we can get rid of $\mathcal{A}.\mathsf{Select}$ from the definition and simply quantify over all worst-case choices of the message m and think of the corresponding aux as being hard-coded in the algorithms $\mathcal{A}.\mathsf{Compress}, \mathcal{A}.\mathsf{Expand}$. However, for adaptive security in the CRS model or security in the random oracle model, we must allow m to be chosen adaptively depending on the CRS or the random oracle queries.

4 HILL-Entropic Encodings and Composition

HILL-Entropic Encodings. We can think of the definition of incompressible encodings as roughly requiring that c has high "Yao incompressibility entropy" [7, 33, 54]. In other words, one cannot efficiently compress c below β bits. We could have defined a stronger variant of "HILL-Entropic Encodings" that would require c to have high HILL entropy [32]. In other words, for any fixed message m, if we consider the random variable C denoting the output of $\mathsf{Enc}(1^\lambda, m)$ then it is computationally indistinguishable from some C' such that the conditional min-entropy $H_\infty(C') \geq \beta$.

Impossibility in the Plain Model. Unfortunately, the notion of "HILL-Entropic Encodings" is simply unachievable in the plain model for any non-trivial $\beta \geq$

$\alpha - k$. In particular, consider the message $m^* = \text{argmin}_m |\mathcal{C}_m|$ that minimizes the size of the set $\mathcal{C}_m = \{c : \mathsf{Dec}(c) = m\}$ of codewords that decode to m. Since $\sum_m |\mathcal{C}_m| \leq 2^\alpha$ we know that $|\mathcal{C}_{m^*}| \leq 2^{\alpha-k}$. Consider a PPT distinguisher $\mathcal{D}$ that gets m^* as non-uniform advice such that $\mathcal{D}(c) = 1$ if $\mathsf{Dec}(c) = m^*$. Then $\Pr[\mathcal{D}(C) = 1] = 1$ for $C \equiv \mathsf{Enc}(1^\lambda, m^*)$ but for any random variable C' such that $H_\infty(C') \geq \alpha - k + 1$ we have

$$\Pr[\mathcal{D}(C') = 1] = \Pr[C' \in \mathcal{C}_{m^*}] = \sum_{c \in \mathcal{C}_{m*}} \Pr[C' = c] \leq 2^{\alpha-k} 2^{-(\alpha-k+1)} \leq \frac{1}{2}.$$

HILL-Entropic Encodings in the CRS Model. On the other, the above impossibility fails in the CRS model and we will construct (selectively-secure) "HILL-Entropic Encodings" in the CRS model. We define these precisely as follows.

Definition 3 (HILL-Entropic Encoding). *An* (α, β)-HILL-Entropic encoding scheme with selective security in the CRS model *consists of PPT algorithms* $(\mathsf{Enc}, \mathsf{Dec})$ *with the same syntax, correctness, and expansion requirements as incompressible encodings. We also require that there is a (potentially inefficient) algorithm* SimEnc. *For any polynomial* $k = k(\lambda)$ *and any ensemble of messages* $m = \{m_\lambda\}$ *of length* $|m_\lambda| = k(\lambda)$*, consider the following "real" experiment:*

- $\mathsf{crs} \leftarrow \{0,1\}^{t(\lambda,k)}$
- $c \leftarrow \mathsf{Enc}_{\mathsf{crs}}(1^\lambda, m_\lambda)$

and let CRS, C *denote the random variables for the corresponding values in the "real" experiment. Also consider the following "simulated" experiment:*

- $(\mathsf{crs}', c') \leftarrow \mathsf{SimEnc}(1^\lambda, m_\lambda)$

and let CRS', C' *denote the random variables for the corresponding values in the "simulated" experiment. We require that* $(\mathsf{CRS}, C) \overset{c}{\approx} (\mathsf{CRS}', C')$ *and* $H_\infty(C' \mid \mathsf{CRS}') \geq \beta(\lambda, k)$.

Theorem 1. *Any* (α, β)*-HILL-Entropic encoding with selective security in the CRS model is also a* (α, β')*-incompressible encoding scheme with selective security in the CRS model, where* $\beta' = \beta - \lambda$.

For lack of space, we defer the proof of the above theorem to the full version.

Composable Encodings in the RO Model. We now show how to convert any selectively HILL-entropic encoding in the CRS model into a composable incompressible encoding scheme in the random oracle model. We rely on a "fixed length" random oracle $\mathsf{RO} : \{0,1\}^\lambda \times \{0,1\}^\lambda \to \{0,1\}^\lambda$. Note that, although we start with a selectively secure scheme in the CRS model where the adversary must choose the message before seeing the CRS, our resulting scheme in the random-oracle model allows the adversary to choose the message after making random-oracle queries.

Our construction proceeds as follows. Assume $(\mathsf{Enc}', \mathsf{Dec}')$ is an encoding in the CRS model with a CRS of length $t(\lambda, k)$. We define:

- $\mathsf{Enc}^{\mathsf{RO}}(1^\lambda, m)$: Choose $r \leftarrow \{0,1\}^\lambda$. Compute $\mathsf{crs} = (\mathsf{RO}(r,1),\ldots,\mathsf{RO}(r,t')) \in \{0,1\}^{t(\lambda,|m|)}$ where $t' = t(\lambda,k)/\lambda$. Let $\hat{c} \leftarrow \mathsf{Enc}'_{\mathsf{crs}}(1^\lambda, m)$. Output $c = (r, \hat{c})$.
- $\mathsf{Dec}^{\mathsf{RO}}(c = (r,\hat{c}))$: Compute $\mathsf{crs} = (\mathsf{RO}(r,1),\ldots,\mathsf{RO}(r,t')) \in \{0,1\}^{t(\lambda,|m|)}$ where $t' = t(\lambda,k)/\lambda$. Output $\mathsf{Dec}'_{\mathsf{crs}}(\hat{c})$.

Theorem 2. *If* $(\mathsf{Enc}', \mathsf{Dec}')$ *is an* (α', β')*-HILL-entropy encoding with selective security in the CRS model, then* $(\mathsf{Enc}, \mathsf{Dec})$ *is a composable* (α, β)*-incompressible encoding in the Random Oracle model where* $\alpha = \alpha' + \lambda$ *and* $\beta = \beta' - \lambda$.

For lack of space, we defer the proof of the above theorem to the full version.

Composability in the Multi-CRS Model. We observe that the above result in the RO model also implies that our CRS model construction is composable in a setting where each encoding is performed with a fresh CRS. In particular, if there were an attack against composable security in the latter setting, it would immediately translate into an attack on the composable security of the former setting.

5 Surjective Lossy Trapdoor Functions

We now introduce our main building block, which we call surjective lossy functions (SLFs). This can be thought of as a relaxation of lossy trapdoor functions [47] that are also *permutations*. In particular, while we insist on the functions being surjective, we relax the requirement that they are injective. Instead, we require a surjective mode and a lossy mode. In surjective mode, we also have an inversion trapdoor. There is some domain distribution D such that, when we sample a random value in the range and invert it using the trapdoor, we get (statistically close to) a random sample from D. In lossy mode, there is some small set of size $\leq 2^\ell$ such that the output of the function almost always ends up in that set - we call ℓ the leakage and want to make it as small as possible.

Definition 4. *A family of* ℓ-surjective lossy trapdoor functions (ℓ-SLFs) *consists of a polynomial-time computable family of functions* $f_{\mathsf{pk}} : \mathcal{D}_{\mathsf{pk}} \to \mathcal{R}_{\mathsf{pk}}$ *along with the following PPT algorithms:*

- $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$*: generates a public-key in lossy mode.*
- $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$*: generates a public-key in surjective mode, along with a trapdoor.*
- $x \leftarrow D(\mathsf{pk})$*: samples a value* $x \in \mathcal{D}_{\mathsf{pk}}$.
- $x \leftarrow \mathsf{Inv}_{\mathsf{td}}(y)$*: samples a pre-image* x *of* $\in \mathcal{R}_{\mathsf{pk}}$.

We require the following properties.

Surjective Mode: *The following distributions over* (pk, x, y) *are statistically indistinguishable:*
  - *Sample* $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$, $x \leftarrow D(\mathsf{pk}), y = f_{\mathsf{pk}}(x)$ *and output* (pk, x, y)

– *Sample* $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$, $y \leftarrow \mathcal{R}_{\mathsf{pk}}, x \leftarrow \mathsf{Inv}_{\mathsf{td}}(y)$ *and output* (pk, x, y).

The above implies that, in particular, we invert correctly:

$$\Pr[f_{\mathsf{pk}}(\mathsf{Inv}_{\mathsf{td}}(y)) = y \ : \ (\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda), y \leftarrow \mathcal{R}_{\mathsf{pk}}] \geq 1 - \mathrm{negl}(\lambda).$$

Lossy Mode: *For any* pk *in the support of* $\mathsf{LossyGen}(1^\lambda)$ *there exists a set* $\mathcal{L}_{\mathsf{pk}}$ *of size* $|\mathcal{L}_{\mathsf{pk}}| \leq 2^{\ell(\lambda)}$ *such that* $\Pr[f_{\mathsf{pk}}(x) \in \mathcal{L}_{\mathsf{pk}} \ : \ \mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda), x \leftarrow D(\mathsf{pk})] = 1 - \mathrm{negl}(\lambda)$.

Indistinguishability: *The following distributions are computationally indistinguishable:*

$$\{\mathsf{pk} : \mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)\} \overset{c}{\approx} \{\mathsf{pk} \ : \ (\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)\}.$$

The above definition captures the main properties of an SLF. However, in the application, we also need some additional properties on the domain, range and the domain distribution. All of the properties would be satisfied ideally if we had a permutation where the domain and range were just $\mathcal{D}_{\mathsf{pk}} = \mathcal{R}_{\mathsf{pk}} = \{0,1\}^{d(\lambda)}$ and the domain distribution $D(\mathsf{pk})$ was just uniform. However, we will need to be more flexible subject to satisfying the following properties.

Definition 5 (SLF*: Enhanced SLF). *We say that an* ℓ*-SLF is an* (r, r', d, e, ℓ)-enhanced SLF*, denoted by SLF*, if the domain* $\mathcal{D}_{\mathsf{pk}}$*, the range* $\mathcal{R}_{\mathsf{pk}}$ *and the domain distribution* $D(\mathsf{pk})$ *satisfy the following properties:*

– *Elements of* $\mathcal{D}_{\mathsf{pk}}$ *can be represented using (at most)* $d(\lambda)$ *bits.*
– *For any fixed* pk*, the min-entropy of the distribution* $D(\mathsf{pk})$ *is at least* $H_\infty(D(\mathsf{pk})) \geq e(\lambda)$.
– *The range* $\mathcal{R}_{\mathsf{pk}}$ *is a group. (We will denote the group operation by addition.)*
– *We can efficiently embed bit-string of length* $r(\lambda)$ *as elements of* $\mathcal{R}_{\mathsf{pk}}$*. In particular, there are efficiently computable functions* $\mathsf{embed}_{\mathsf{pk}} : \{0,1\}^{r(\lambda)} \rightarrow \mathcal{R}_{\mathsf{pk}}$ *and* $\mathsf{unembed}_{\mathsf{pk}} : \mathcal{R}_{\mathsf{pk}} \rightarrow \{0,1\}^{r(\lambda)}$ *such that for all* $m \in \{0,1\}^{r(\lambda)}$ *we have*

$$\Pr[\mathsf{unembed}_{\mathsf{pk}}(\mathsf{embed}_{\mathsf{pk}}(m)) = m \ : \ (\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)] = 1 - \mathrm{negl}(\lambda).$$

– *We can obliviously sample uniformly random elements of* $\mathcal{R}_{\mathsf{pk}}$*. In particular, there exists some PPT algorithm* $y \leftarrow \mathsf{sam}(\mathsf{pk})$ *that uses* $r'(\lambda)$ *random bits to sample a uniformly random values* $y \in \mathcal{R}_{\mathsf{pk}}$ *along with a PPT algorithm* $\mathsf{explain}_{\mathsf{pk}}(y)$ *such that* $(u, y) \overset{s}{\approx} (u', y')$*, where* $u \leftarrow \{0,1\}^{r'(\lambda)}$*,* $y = \mathsf{sam}(\mathsf{pk}; u)$*,* $y' \leftarrow \mathcal{R}_{\mathsf{pk}}$*,* $u' \leftarrow \mathsf{explain}_{\mathsf{pk}}(y)$.

We say that a scheme is a "good" SLF, without specifying parameters, if it is an* (r, r', d, e, ℓ)*-SLF* for some* $r = r(\lambda)$ *with* $d = (1 + o(1))r$*,* $r' = (1 + o(1))r$*,* $e = (1 - o(1))r$*, and* $\ell = o(r)$.

5.1 SLFs from Decision Composite Residuosity

We describe a construction of a good SLF* under the Decision Composite Residuosity (DCR) assumption of Paillier [44]. The construction is identical to that of [25] and is based on the Damgård-Jurik Cryptosystem [18]. We provide it here for completeness.

The Damgård-Jurik Cryptosystem. Let $N = PQ$ where P, Q are odd primes such that $\gcd(N, \varphi(N)) = 1$. We call such N admissible. When P and Q are sufficiently large and randomly chosen,$N = PQ$ is admissible with all but negligible probability. The following theorem gives the structure of the group $\mathbb{Z}^*_{N^{s+1}}$.

Theorem 3 ([18]). *For any admissible $N = PQ$ and $s < \min\{P, Q\}$ the map $\psi_{N,s} \ : \ \mathbb{Z}_{N^s} \times \mathbb{Z}^*_N \to \mathbb{Z}^*_{N^{s+1}}$ given by $\psi_{N,s}(m, r) = (1 + N)^m r^{N^s} \mod N^{s+1}$ is an isomprphism satisfying*

$$\psi_{N,s}(m_1 + m_2 \bmod N, r_1 r_2 \bmod N^s) = \psi_{N,s}(m_1, r_1) \cdot \psi_{N,s}(m_2, r_2) \bmod N^{s+1}.$$

Moreover, $\psi_{N,s}$ can be inverted in polynomial time given P, Q.

In the Damgård-Jurik [18] cryptosystem, the public key is N and the secret key is P, Q. The encryption of a message $m \in \mathbb{Z}_{N^s}$ is $\psi_{N,s}(m, r)$ for a random $r \in \mathbb{Z}^*_N$ and the decryption inverts $\psi_{N,s}$ using the secret key. The cryptosystem is proven secure under the decision composite residuosity (DCR) assumption stated below.

Definition 6 ([44]). *The decision composite residuosity (DCR) assumption states that for randomly chosen primes P, Q in the range $[2^{\lambda-1}, 2^\lambda]$ and $N = PQ$ the distributions (N, x) and (N, y) are computatinally indistinguishable where $x \leftarrow \mathbb{Z}^*_{N^2}$ is uniformly random and $y \leftarrow \{z^N \bmod N^2 \ : \ z \in \mathbb{Z}^*_N\}$ is a random N-residue over $\mathbb{Z}^*_{N^2}$.*

Intuitively the DCR assumption states that for $s = 1$, one cannot distinguish between $\psi_{N,s}(m, r)$ versus $\psi_{N,s}(0, r)$ for a uniformly random m, r. It's easy to see that this implies that for any fixed m, m' one cannot distinguish between $\psi_{N,s}(m, r)$ versus $\psi_{N,s}(m', r)$. Moreover, it turns out that the DRC assumption, which is stated for $s = 1$, automatically implies security for arbitrary polynomial s. The following theorem essentially states the that the Damgård-Jurik cryptosystem is semantically secure under the DCR assumption.

Theorem 4 ([18]). *For any polynomial $s = \mathrm{poly}(\lambda)$, and for randomly chosen primes P, Q in the range $[2^{\lambda-1}, 2^\lambda]$ with $N = PQ$ and for any values $m, m' \in \mathbb{Z}_{N^s}$, the distributions $(N, \psi_{N,s}(m, r))$ and $(N, \psi_{N,s}(m', r))$ over $r \leftarrow \mathbb{Z}^*_N$ are computationally indistinguishable under the DCR assumption.*

Constructing SLFs from DCR. Since $\psi_{N,s}$ is a permutation with cryptographic properties, we could think of setting $\psi_{N,s}$ as the SLF. Unfortunately, it's not clear how to make it lossy directly. Instead, in addition to the modulus N, we add a "ciphertext" $c \in \mathbb{Z}^*_{N^{s+1}}$ to the public key and define

$$f_{\mathsf{pk}} \ : \ \mathbb{Z}_{N^s} \times \mathbb{Z}^*_N \to \mathbb{Z}^*_{N^{s+1}} \quad \text{given by} \quad f_{\mathsf{pk}}(x = (m, r)) = c^m \psi_{N,s}(0, r)$$

In surjective (bijective) mode, we set $c = \psi_{N,s}(1, \hat{r})$ to be an encryption of the message 1, and we also add the randomness $\hat{r}$ to the secret key. In that case, $f_{\mathsf{pk}}(m, r) = c^m \psi_{N,s}(0, r) = \psi_{N,s}(m, \hat{r}^m \cdot r)$. We can invert f_{pk} on any value $y \in \mathbb{Z}^*_{N^{s+1}}$ using the secret key, by computing $\psi^{-1}_{N,s}(y) = (m, r')$ and outputting $x = (m, r)$ were $r = r'/\hat{r}^m$. In lossy mode, we set $c = \psi_{N,s}(0, \hat{r})$ to be a random encryption of 0. In that case, the image of the function f_{pk} is the set $\mathcal{L}_{\mathsf{pk}} = \{\psi_{N,s}(0, r') \ : \ r' \in \mathbb{Z}^*_N\}$. In other words, in lossy mode, $f_{\mathsf{pk}}(x)$ only contains $\approx \log(N)$ bits of information about the $\approx (s+1)\log N$ bit value x.

We describe the construction in detail below. Let $s = s(\lambda)$ be a parameter.

- $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$: Generate random λ-bit primes P, Q such that $N = PQ$ is admissible. Let $\hat{r} \leftarrow \mathbb{Z}^*_N$ and set $c = \psi_{N,s}(1, \hat{r})$. Output $\mathsf{pk} = (N, c), \mathsf{sk} = (P, Q, \hat{r})$.
- $\mathsf{LossyGen}(1^\lambda)$: Generate random λ-bit primes P, Q such that $N = PQ$ is admissible. Let $\hat{r} \leftarrow \mathbb{Z}^*_N$ and set $c = \psi_{N,s}(0, \hat{r})$. Output $\mathsf{pk} = (N, c)$.
- $y = f_{\mathsf{pk}}(x)$: The function $f_{\mathsf{pk}} \ : \ \mathcal{D}_{\mathsf{pk}} \to \mathcal{R}_{\mathsf{pk}}$ has domain $\mathcal{D}_{\mathsf{pk}} = \mathbb{Z}_{N^s} \times \mathbb{Z}^*_N$ and range $\mathcal{R}_{\mathsf{pk}} = \mathbb{Z}^*_{N^{s+1}}$. It is defined by $f_{\mathsf{pk}}(x) = c^m \psi_{N,s}(0, r) \bmod N^{s+1}$, where $x = (m, r) \in \mathbb{Z}_{N^s} \times \mathbb{Z}^*_N$.
- $x = (m, r) \leftarrow D(\mathsf{pk})$: samples a uniformly random value $x \leftarrow \mathcal{D}_{\mathsf{pk}}$.
- $x \leftarrow \mathsf{Inv}_{\mathsf{td}}(y)$: use the secret key P, Q to computr $\psi^{-1}_{N,s}(y) = (m, r')$ and output $x = (m, r)$ were $r = r'/\hat{r}^m$.

Theorem 5. *For any polynomial $s = s(\lambda)$ the above construction is an (r, r', d, e, ℓ)-SLF* where:*

$$r = (s+1)2(\lambda - 1), \quad r' = (s+1)2\lambda + \lambda, \quad d = (s+1)2\lambda, \quad e = (s+1)2(\lambda-1) - 1, \quad \ell = 2\lambda$$

In particular, when $s = \omega(1)$ then the above construction is a good SLF.*

Proof. We begin by showing each of the SLF properties:

- Surjective Mode: When $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$ is sampled in surjective mode, the function f_{pk} is a bijection over $\mathcal{D}_{\mathsf{pk}} \cong \mathcal{R}_{\mathsf{pk}}$ and $\mathsf{Inv}_{\mathsf{td}}$ is the inverse of f_{pk}. In particular, the distribution of (pk, x, y) for $x \leftarrow D(\mathsf{pk})$, $y = f_{\mathsf{pk}}(x)$ is identical to sampling $y \leftarrow \mathcal{R}_{\mathsf{pk}}$ and $x = \mathsf{Inv}_{\mathsf{td}}(y)$.
- Lossy Mode: When $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$ is sampled in lossy mode, we have
$$\mathcal{L}_{\mathsf{pk}} = \{f_{\mathsf{pk}}(x) \ : \ x \in \mathcal{D}_{\mathsf{pk}}\} = \{\psi_{N,s}(0, r') \ : \ r' \in \mathbb{Z}^*_N\}$$
and therefore $|\mathcal{L}_{\mathsf{pk}}| \le \log N \le 2\lambda$.
- Indistinguishability: This follows immediately from Theorem 4 with $m = 0$ and $m' = 1$.

Next we discuss the augmented properties to show that the above SLF is also an SLF*.

- Elements of $\mathcal{D}_{\mathsf{pk}} = \mathbb{Z}_{N^s} \times \mathbb{Z}^*_N$ can be represented using $d(\lambda) = (s+1)2\lambda$ bits.
- The min-entropy of the distribution $D(\mathsf{pk})$, which is uniform over $\mathcal{D}_{\mathsf{pk}}$ is $\log|\mathcal{D}_{\mathsf{pk}}| \ge (s+1)2(\lambda - 1) - 1$. This is because $N \ge 2^{2(\lambda-1)}$.
- The range $\mathcal{R}_{\mathsf{pk}} = \mathbb{Z}^*_{N^{s+1}}$ is a group under multiplication.

– We can efficiently embed bit-string of length $r(\lambda) = (s+1)2(\lambda-1)-1$ as elements of $\mathcal{R}_{\mathsf{pk}}$. We do so, by simply taking the string and interpreting it as an integer $y < N^{s+1}$ in binary. The probability of $y \notin \mathbb{Z}^*_{N^{s+1}}$ is negligible over the random choice of $N = PQ$.
– We can obliviously sample from $\mathcal{R}_{\mathsf{pk}} = \mathbb{Z}^*_{N^{s+1}}$ using $r'(\lambda) = ((s+1)2\lambda + \lambda)$-bits. We do so by defining $\mathsf{sam}(\mathsf{pk})$ to choose a random $r'(\lambda)$-bit integer z and outputting $y = z \bmod N^{s+1}$. This is $2^{-\lambda}$ statistically close to sampling $y \leftarrow \mathbb{Z}_N^{s+1}$ which is statistically close to sampling $y \leftarrow \mathbb{Z}^*_{N^{s+1}}$. The $\mathsf{explain}_{\mathsf{pk}}(y)$ algorithm outputs a random $r'(\lambda)$-bit value z such that $z = y \bmod N^{s+1}$; it does so by setting $t = \lfloor 2^{r'}/N^{s+1} \rfloor$ and outputting $z = y + v \cdot N^{s+1}$ where $v \leftarrow \{0, \ldots, t\}$. For any y, $z = \mathsf{explain}(y)$ is uniformly random over all z such that $\mathsf{sam}(\mathsf{pk}; z) = y$.

5.2 SLFs from Learning with Errors

Lattice Preliminaries For any integer $q \geq 2$, we let $\mathbb{Z}_q$ denote the ring of integers modulo q. For a vector $\mathbf{e} \in \mathbb{Z}^n$ we write $||\mathbf{e}||_\infty \leq \beta$ if each entry e_i in $\mathbf{e}$ satisfies $|e_i| \leq \beta$. Similarly, for a matrix $\mathbf{E} \in \mathbb{Z}_q^{n\times m}$ we write $||\mathbf{E}||_\infty \leq \beta$ if each entry $e_{i,j}$ in $\mathbf{E}$ satisfies $|e_{i,j}| \leq \beta$. We say that a distribution χ over $\mathbb{Z}$ is β-bounded if $\Pr[|x| \leq \beta \;:\; x \leftarrow \chi] \leq \mathrm{negl}(\lambda)$. By default, all vectors are assumed to be *column* vectors. For integers $q \geq p \geq 2$ we define the rounding function

$$\lceil \cdot \rfloor_p \;:\; \mathbb{Z}_q \to \mathbb{Z}_p \;:\; x \mapsto \lfloor (p/q)\cdot x \rceil$$

If p divides q then the rounding function divides $\mathbb{Z}_q$ into p intervals of size q/p each. If $q = 2^k$ and $p = 2^{k'}$ then $\lceil x \rfloor_p$ corresponds to outputting the k' most significant bits of the k-bit integer x. If $\mathbf{x}$ is a vector we let $\lceil \mathbf{x} \rfloor_p$ denote the component-wise rounding operation.

Learning with Errors (LWE). The learning with errors (LWE) assumption was introduced by Regev in [48].

Definition 7 ([48]). *Let n, q be integers and χ a probability distribution over $\mathbb{Z}_q$, all parameterized by the security parameter λ. The (n, q, χ)-LWE assumption says that for all polynomial m the following distributions are computationally indistinguishable*

$$(\mathbf{A}, \mathbf{A}\mathbf{s} + \mathbf{e}) \overset{c}{\approx} (\mathbf{A}, \mathbf{u}^t) \qquad : \mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}, \mathbf{s} \leftarrow \mathbb{Z}_q^n, \mathbf{e} \leftarrow \chi^m, \mathbf{u} \leftarrow \mathbb{Z}_q^m.$$

The work of [5] showed that the (n, q, χ)-LWE assumption above also implies security when the secret is chosen from the error distribution χ:

$$(\mathbf{A}, \mathbf{A}\mathbf{s} + \mathbf{e}) \overset{c}{\approx} (\mathbf{A}, \mathbf{u}) \qquad : \mathbf{A} \leftarrow \mathbb{Z}_q^{m\times n}, \mathbf{s} \leftarrow \chi^n, \mathbf{e} \leftarrow \chi^m, \mathbf{u} \leftarrow \mathbb{Z}_q^m.$$

The works of [10,45,48] show that the LWE assumption is as hard as (quantum) solving GapSVP and SIVP under various parameter regimes. In particular, we will assume that for every $q = 2^{\mathrm{poly}(\lambda)}$ there exists some polynomial

$n = \text{poly}(\lambda)$ and $\beta = \text{poly}(\lambda)$ along with β-bounded distribution χ such that the $\mathsf{LWE}_{n,q,\chi}$ assumption holds. We refer to the above as the LWE assumption when we don't specify parameters. This is known to be as hard as solving GapSVP and (quantum) SIVP with sub-exponential approximation factors, which is believed to be hard.

The Gadget Matrix and Preimage Sampling. Let $q = B^\gamma$ be a modulus. We define the base-B gadget matrix of dimension n as the matrix $\mathbf{G} \in \mathbb{Z}_q^{n \times \gamma n}$ given by

$$\mathbf{G} = \mathbf{I}_n \times \mathbf{g} = \begin{bmatrix} \cdots \mathbf{g}^t \cdots & & & \\ & \cdots \mathbf{g}^t \cdots & & \\ & & \ddots & \\ & & & \cdots \mathbf{g}^t \cdots \end{bmatrix}$$

where $\mathbf{g} = [1, B, B^2, \ldots, B^{\gamma-1}]$.

Lemma 2 (Preimage Sampling [27, 42]). *Let $q = B^\gamma$ and n, n' be some parameters such that $n, n', \log q$ are polynomial in the security parameter λ and $n \geq \lambda$. There exist PPT algorithms $\mathsf{SamPre}, \mathsf{Sam}$ such that the following holds. Let $\mathbf{G} \in \mathbb{Z}_q^{n \times \gamma n}$ be the base-B gadget matrix of dimension n. Let $\overline{\mathbf{A}} \in \mathbb{Z}_q^{n \times n'}$ and let $\mathbf{A} = [\overline{\mathbf{A}} \mid \overline{\mathbf{A}}\mathbf{R} + \mathbf{G}] \in \mathbb{Z}_q^{n \times m}$ where $m = n' + \gamma n$ and $\mathbf{R} \in \mathbb{Z}_q^{n' \times \gamma n}$ with $||\mathbf{R}||_\infty \leq \beta$. Then:*

– $\mathbf{u} \leftarrow \mathsf{Sam}(1^\lambda)$ *samples* $\mathbf{u} \in \mathbb{Z}_q^m$ *such that* $||\mathbf{u}||_\infty \leq m^{2.5}\beta B$,
– $\mathbf{u} \leftarrow \mathsf{SamPre}_{\mathbf{A},\mathbf{R}}(\mathbf{v})$ *samples* $\mathbf{u} \in \mathbb{Z}_q^m$ *such that* $\mathbf{A}\mathbf{u} = \mathbf{v}$ *and* $||\mathbf{u}||_\infty \leq m^{2.5}\beta B$,

where the distribution of $(\mathbf{u}, \mathbf{v})$ is statistically close to $(\mathbf{u}', \mathbf{v}')$ with $\mathbf{u} \leftarrow \mathsf{Sam}(1^\lambda), \mathbf{v} = \mathbf{A}\mathbf{u}$, $\mathbf{v}' \leftarrow \mathbb{Z}_q^n, \mathbf{u}' \leftarrow \mathsf{SamPre}_{\mathbf{A},\mathbf{R}}(\mathbf{v}')$. Furthermore, the distribution of $\mathsf{Sam}(1^\lambda)$ has min-entropy $H_\infty(\mathsf{Sam}(1^\lambda)) \geq m \log B$.

The above lemma follows from the works of [27, 42]. Firstly, [42] shows that the lattice $\Lambda^\perp(\mathbf{G})$ has a public basis $\mathbf{S} \in \mathbb{Z}^{m' \times m'}$ with $||\mathbf{S}||_\infty \leq B$ where $m' = \gamma n$. Furthermore, this can be efficiently extended to a basis $\mathbf{T} \in \mathbb{Z}^{m \times m}$ for the lattice $\Lambda^\perp(\mathbf{A})$ using knowledge of $\mathbf{A}, \mathbf{R}$, where $||\mathbf{T}||_\infty \leq m'\beta B$. This also shows that the smoothing parameter of $\eta_\varepsilon(\Lambda^\perp(\mathbf{A})) \leq ||\mathbf{T}|| \leq m^{1.5}\beta B$. We define $\mathsf{Sam}(1^\lambda)$ to sample from the Discrete Gaussian $D_{\mathbb{Z}^m,s}$ with parameter $s = m^2\beta B$. Following [27], the algorithm $\mathsf{SamPre}_{\mathbf{A},\mathbf{R}}(\mathbf{v})$ uses $\mathbf{A}, \mathbf{R}$ to find the basis $\mathbf{T}$ and uses that to sample from the Discrete Gaussaian $\mathbf{t} + D_{\Lambda^\perp(\mathbf{A}),s,-\mathbf{t}}$ where $\mathbf{t}$ is any solution to $\mathbf{A}\mathbf{t} = \mathbf{v}$ (which is guaranteed to exist and can be found efficiently). For $\mathbf{u} \leftarrow \mathsf{Sam}(1^\lambda)$ the value $\mathbf{A}\mathbf{u}$ is then statistically close to uniform over $\mathbb{Z}_q^n$ and for any $\mathbf{v}$, the distribution of $\mathbf{u}' \leftarrow \mathsf{SamPre}_{\mathbf{A},\mathbf{R}}(\mathbf{v})$ is exactly the conditional distribution of $\mathbf{u} \leftarrow \mathsf{Sam}(1^\lambda)$ subject to $\mathbf{A}\mathbf{u} = \mathbf{v}$. The probability that $\mathbf{u} \leftarrow \mathsf{Sam}(1^\lambda)$ or $\mathbf{u}' \leftarrow \mathsf{SamPre}_{\mathbf{A},\mathbf{R}}(\mathbf{v})$ have norm greater than $s\sqrt{m} = m^{2.5}\beta B$ is negligible and, for simplicity, we will modify the algorithms to never output such values. Lastly, we rely on Lemma 2.11 in [46] and the fact that the min-entropy of the discrete Gaussian $D_{\mathbb{Z}^m,s}$ is greater than $(s/(2\eta_\varepsilon(\mathbb{Z}^m)))^m \geq (s/\log m)^m \geq B^m$ for the min-entropy claim.

Constructing of SLFs from LWE Let $n > n'$ and $B, q = B^\gamma, C, p$ be parameters depending on the security parameter λ and assume that p divides q. Let χ be some β-bounded error distribution. Define $m' = \gamma n$ and $m = n' + m'$. Let $\mathbf{G}$ be the base-B gadget matrix of dimension n over $\mathbb{Z}_q$.

For a public key $\mathsf{pk} = \mathbf{A} \in \mathbb{Z}_q^{n\times m}$ define the function $f_{\mathsf{pk}} : \{-C, \dots, C\}^m \to \mathbb{Z}_p^n$ via

$$f_{\mathsf{pk}}(\mathbf{x}) = \lceil \mathbf{A}\mathbf{x} \rfloor_p .$$

The domain is $\mathcal{D} = \{-C, \dots, C\}^m$ and the range is $\mathcal{R} = \mathbb{Z}_p^n$. We define all of the algorithms of the SLF as follows

- $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$: Choose a random $\overline{\mathbf{A}} \leftarrow \mathbb{Z}_q^{n\times n'}$, $\mathbf{R} \leftarrow \chi^{n'\times m'}, \mathbf{E} \leftarrow \chi^{n\times m'}$ and set

$$\begin{aligned} \mathsf{pk} &= \mathbf{A} = [\overline{\mathbf{A}} \mid \overline{\mathbf{A}}\mathbf{R} + \mathbf{G} + \mathbf{E}] \\ \mathsf{td} &= (\mathbf{A}, \mathbf{R}, \mathbf{E}) \end{aligned}$$

- $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$: Choose a random $\overline{\mathbf{A}} \leftarrow \mathbb{Z}_q^{n\times n'}$, $\mathbf{S} \leftarrow \mathbb{Z}_q^{n'\times m}$, $\mathbf{E} \leftarrow \chi^{n\times m}$ and set

$$\mathbf{A} = \overline{\mathbf{A}} \cdot \mathbf{S} + \mathbf{E}$$

- $\mathbf{x} \leftarrow D(\mathsf{pk})$: samples a uniformly random value $\mathbf{x} \leftarrow \mathsf{Sam}(1^\lambda)$.
- $\mathbf{x} \leftarrow \mathsf{Inv}_{\mathsf{td}}(\mathbf{y})$: Choose $\mathbf{v} \in \mathbb{Z}_q^n$ uniformly at random subject to $\lceil \mathbf{v} \rfloor_p = \mathbf{y}$ by choosing each coordinate uniformly from the appropriate interval. Let $\mathbf{A}' = \mathbf{A} - [\mathbf{0}|\mathbf{E}] = [\overline{\mathbf{A}} \mid \overline{\mathbf{A}}\mathbf{R} + \mathbf{G}]$ and output $\mathbf{x} \leftarrow \mathsf{SamPre}_{\mathbf{A}',\mathbf{R}}(\mathbf{v})$.

For concreteness, we set $B = 2^\lambda$, we set $\gamma = \lambda$ so that $q = B^\gamma = 2^{\lambda^2}$ and $p = 2^{\lambda^2 - 2\lambda}$. By the LWE assumption, there are some polynomials $n' = \mathrm{poly}(\lambda), \beta = \mathrm{poly}(\lambda)$ and a β-bounded distribution χ so that the $\mathsf{LWE}_{n',q,\chi}$ assumption hold. We set $n = n' \cdot \lambda$. Lastly, we choose $C = \lceil m^{2.5}\beta B \rceil$. This ensures that $\mathbf{x} \leftarrow \mathsf{Sam}(1^\lambda)$ outputs $\mathbf{x}$ such that $||\mathbf{x}||_\infty \le C$.

Theorem 6. *The above construction is an (r, r', d, e, ℓ)-SLF* under the LWE assumption where*

$$\begin{aligned} &r = r' = n \log p = n(\lambda^2 - 2\lambda) = \lambda^2 n(1 - o(1)) \\ &d = \lceil m \log(2C + 1) \rceil = (n + \gamma n)(b + O(\log m + \log \beta) = \lambda^2 n(1 + o(1)) = (1 + o(1))r \\ &e = m \log B = (n + \gamma n)\lambda \ge \lambda^2 n \ge r \\ &\ell = n' \log q \le (n/\lambda)\lambda^2 = o(r) \end{aligned}$$

In particular, it is a good SLF.*

Proof. We show each property of SLFs in turn below:

- Surjective Mode: Let $(\mathsf{pk} = \mathbf{A}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$ be some key pair sampled in surjective mode with $\mathbf{A} = [\overline{\mathbf{A}} \mid \overline{\mathbf{A}}\mathbf{R} + \mathbf{G} + \mathbf{E}]$ and let $\mathbf{A}' = \mathbf{A} - [\ \mathbf{0} \mid \mathbf{E}\ ] = [\overline{\mathbf{A}} \mid \overline{\mathbf{A}}\mathbf{R} + \mathbf{G}]$. By Lemma 2, we have

$$(\mathbf{x}, \mathbf{A}'\mathbf{x}) \overset{\mathrm{s}}{\approx} (\mathbf{x}', \mathbf{v})$$

where $\mathbf{x} \leftarrow \mathsf{Sam}(1^\lambda)$, $\mathbf{v} \leftarrow \mathbb{Z}_q^n$ and $\mathbf{x}' \leftarrow \mathsf{SamPre}_{\mathbf{A}',\mathbf{R}}(\mathbf{v})$. This implies that

$$(\mathbf{x}, \lceil \mathbf{A}\mathbf{x} \rfloor_p) \equiv (\mathbf{x}, \lceil \mathbf{A}'\mathbf{x} + [\ \mathbf{0} \mid \mathbf{E}\ ]\mathbf{x} \rfloor_p) \overset{\text{s}}{\approx} (\mathbf{x}', \lceil \mathbf{v} + [\ \mathbf{0} \mid \mathbf{E}\ ]\mathbf{x}' \rfloor_p) \overset{\text{s}}{\approx} (\mathbf{x}', \lceil \mathbf{v} \rfloor_p) \equiv (\mathbf{x}', \mathbf{y})$$

where $\mathbf{y} \leftarrow \mathbb{Z}_p^n$. Here we use the fact that $\Pr[\lceil \mathbf{v} \rfloor_p \neq \lceil \mathbf{v} + \mathbf{e} \rfloor_p] = \text{negl}(\lambda)$ where $\mathbf{e} = [\ \mathbf{0} \mid \mathbf{E}\ ]\mathbf{x}'$. This is because $\mathbf{v}$ is uniform over $\mathbb{Z}_q$ and $\mathbf{e}$ has norm $\tau = ||\mathbf{e}||_\infty \leq Cm'\beta = 2^{\lambda + O(\log \lambda)}$. Therefore the only way that $\lceil \mathbf{v} + \mathbf{e} \rfloor_p \neq \lceil \mathbf{v} \rfloor_p$ is if some coordinate of $\mathbf{v}$ lies within distance τ of the boundary of an interval of size q/p that gets rounded to the same value, but for any coordinate this happens with probability $(2\tau + 1)/(q/p) = 2^{\lambda + O(\log \lambda)}/2^{2\lambda} = \text{negl}(\lambda)$.

– Lossy Mode: For $(\mathsf{pk} = \mathbf{A}) \leftarrow \mathsf{LossyGen}(1^\lambda)$ we have $\mathbf{A} = \overline{\mathbf{A}}\mathbf{S} + \mathbf{E}$. We define

$$\mathcal{L}_{\mathsf{pk}} = \left\{ \lceil \overline{\mathbf{A}}\mathbf{S}\mathbf{x} \rfloor_p \ : \ \mathbf{x} \in \{-C, \dots, C\}^m \right\} \subseteq \left\{ \lceil \overline{\mathbf{A}}\mathbf{y} \rfloor_p \ : \ \mathbf{y} \in \mathbb{Z}_q^{n'} \right\}$$

which is of size $|\mathcal{L}_{\mathsf{pk}}| \leq q^{n'} \leq 2^\ell$. For any $\mathbf{x} \in \{-C, \dots, C\}^m = \mathcal{D}_{\mathsf{pk}}$ and a random $\mathbf{A} \leftarrow \mathsf{LossyGen}(1^\lambda)$ we have

$$\Pr\left[\lceil \mathbf{A}\mathbf{x} \rfloor_p \notin \mathcal{L}_{\mathsf{pk}}\right] \leq \Pr\left[\lceil \overline{\mathbf{A}}\mathbf{S}\mathbf{x} + \mathbf{E}\mathbf{x} \rfloor_p \neq \lceil \overline{\mathbf{A}}\mathbf{S}\mathbf{x} \rfloor_p\right] \leq \text{negl}(\lambda)$$

Here, we rely on the fact that $\tau = ||\mathbf{E}\mathbf{x}||_\infty \leq Cm\beta = 2^{\lambda + O(\log \lambda)}$. If $\mathbf{S}\mathbf{x} = \mathbf{0}$ then the above can never happen. Otherwise $\overline{\mathbf{A}}\mathbf{S}\mathbf{x}$ is uniformly random over the choice of $\overline{\mathbf{A}}$ and the above can only happen if some some coordinate of $\overline{\mathbf{A}}\mathbf{S}\mathbf{x}$ lies within distance τ of the boundary of an interval of size q/p that gets rounded to the same value, but for any coordinate this happens with probability $(2\tau + 1)/(q/p) = 2^{\lambda + O(\log \lambda)}/2^{2\lambda} = \text{negl}(\lambda)$.

– Indistinguishability: We claim that the distributions of pk sampled from either $\mathsf{LossyGen}(1^\lambda)$ or $\mathsf{SurGen}(1^\lambda)$ are both computationally indistinguishable from the uniform distribution over $\mathbb{Z}_q^{n \times m}$, and therefore also indistinguishable from each other.
Firstly for $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$ we have $\mathsf{pk} = [\overline{\mathbf{A}} \mid \overline{\mathbf{A}}\mathbf{R} + \mathbf{G} + \mathbf{E}]$. By thinking of the columns of $\mathbf{R}$ as LWE secrets that come from the error distribution and $\overline{\mathbf{A}}$ as the LWE coefficients, we get that $[\overline{\mathbf{A}}, \overline{\mathbf{A}}\mathbf{R} + \mathbf{E}]$ is computationally indistinguishable from uniform, which also shows that $\mathsf{pk} = [\overline{\mathbf{A}}, \overline{\mathbf{A}}\mathbf{R} + \mathbf{E}] + [\mathbf{0} \mid \mathbf{G}]$ is indistinguishable from uniform.
Second for $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$ we have $\mathsf{pk} = [\overline{\mathbf{A}}\mathbf{S} + \mathbf{E}]$. By thinking of the columns of $\mathbf{S}$ as LWE secrets and $\overline{\mathbf{A}}$ as the LWE coefficients, it is immediate that pk is computationally indistinguishable from uniform.

Next, we prove that the domain and range satisfy the enhanced properties that make it an SLF*.

– Elements of $\mathcal{D}_{\mathsf{pk}} = \{-C, \dots, C\}^m$ can be represented using $d = \lceil m \log(2C + 1) \rceil$ bits.
– By Lemma 2, the min-entropy of the distribution $D(\mathsf{pk}) = \mathsf{Sam}(1^\lambda)$ is at least $m \log B$.
– The range $\mathcal{R}_{\mathsf{pk}} = \mathbb{Z}_p^n$ is a group under addition (or we can interpret $\mathcal{R}_{\mathsf{pk}} = \{0,1\}^{n(\lambda^2 - 2\lambda)}$ as a group under XOR).

- We can efficiently embed bit-string of length $r = n \log p = n(\lambda^2 - 2\lambda)$ as elements of $\mathcal{R}_{\mathsf{pk}} = \mathbb{Z}_p^n$.
- We can obliviously sample from $\mathcal{R}_{\mathsf{pk}} = \mathbb{Z}_p^n$ using r' random bits by equating elements of $\mathbb{Z}_p^n$ with bit-strings of length $r' = n \log p$.

6 Incompressible Encodings from SLFs

We now construct incompressible encodings in the CRS model. We will show that these encodings satisfy the stronger notion of HILL-entropic security (Definition 3 from Section 4), which implies incompressibility by Theorem 1. Furthermore, this means that we can also use this construction to get a composable incompressible encoding in the random oracle model using Theorem 2.

Construction. Given an (r, r', d, e, ℓ)-SLF*, we define the incompressible encoding scheme in the CRS model as follows:

- $\mathsf{crs} = (u_1, \ldots, u_{k'}) \leftarrow \{0,1\}^{r'(\lambda)\cdot k'}$, where $r'(\lambda)$ is the number of random bits needed to obliviously sample from the range of the SLF and $k' = \lceil k/r(\lambda) \rceil$.
- $\mathsf{Enc}_{\mathsf{crs}}(1^\lambda, m)$: Choose $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$. Interpret $m = (m_1, \ldots, m_{k'}) \in \{0,1\}^{r(\lambda)\cdot k'}$. For $i \in [k']$, let $\hat{m}_i = \mathsf{embed}_{\mathsf{pk}}(m_i) \in \mathcal{R}_{\mathsf{pk}}$, $y_i := \mathsf{sam}(\mathsf{pk}; u_i) \in \mathcal{R}_{\mathsf{pk}}$ and $c_i \leftarrow \mathsf{Inv}_{\mathsf{td}}(y_i + \hat{m}_i)$. Output $c = (\mathsf{pk}, c_1, \ldots, c_{k'})$.
- $\mathsf{Dec}_{\mathsf{crs}}(\ c = (\mathsf{pk}, c_1, \ldots, c_{k'}))$: For $i \in [k']$, let $y_i := \mathsf{sam}(\mathsf{pk}; u_i)$, $\hat{m}_i := f_{\mathsf{pk}}(c_i) - y_i$, $m_i := \mathsf{embed}_{\mathsf{pk}}^{-1}(\hat{m}_i)$. Output $(m_1, \ldots, m_{k'})$.

Theorem 7. *Assuming the existence of an (r, r', d, e, ℓ)-SLF*, the above construction yields an (α, β)-HILL-Entropic encoding scheme with selective security in the CRS model, where $\alpha(\lambda, k) = k'(\lambda)d(\lambda) + \mathrm{poly}(\lambda)$, $\beta(\lambda, k) = k'(\lambda)(e(\lambda) - \ell(\lambda))$ and the crs is of length $t(\lambda, k) = k'(\lambda) \cdot r'(\lambda)$ for $k'(\lambda) = \lceil \frac{k}{r(\lambda)} \rceil$. Furthermore, the encoding is locally decodable.*

In particular, any good SLF yields a good incompressible encoding that is locally decodable and achieves either:*

1. *Selective security in the CRS model, where the CRS is of length $t(\lambda, k) = k(1 + o(1)) + \mathrm{poly}(\lambda)$.*
2. *Composable security in the Random Oracle model.*

Proof. We only prove the first part of the theorem and the second part ("in particular...") follows from Theorems 1 and 2.

The correctness of the scheme and the parameter α (length of encoding), t (length of CRS) are clear from the construction. To show β-HILL-Entropic security, define the procedure $(\mathsf{crs}, c) \leftarrow \mathsf{SimEnc}(1^\lambda, m)$ that, on input $m = (m_1, \ldots, m_{k'}) \in \{0,1\}^{r(\lambda)\cdot k'}$, samples $\mathsf{crs} = (u_1, \ldots, u_{k'})$ and $c = (\mathsf{pk}, c_1, \ldots, c_{k'})$ as follows:

- Choose $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$.
- For $i \in [k']$, choose $c_i \leftarrow D(\mathsf{pk})$.

– For $i \in [k']$, let $\hat{m}_i = \mathsf{embed}_{\mathsf{pk}}(m_i) \in \mathcal{R}_{\mathsf{pk}}$. Let $y_i = f_{\mathsf{pk}}(c_i) - \hat{m}_i$ if $f_{\mathsf{pk}}(c_i) \in \mathcal{L}_{\mathsf{pk}}$ or else set $y_i = L - \hat{m}_i$ where L is some arbitrary element of $\mathcal{L}_{\mathsf{pk}}$. Let $u_i \leftarrow \mathsf{explain}_{\mathsf{pk}}(y_i)$.

First we show that SimEnc satisfies the entropy requirements. For any fixed m, pk, let $\mathsf{CRS} = (U_1, \ldots, U_{k'}), C = (\mathsf{pk}, C_1, \ldots, C_{k'})$ be random variables for the output $(\mathsf{crs}, c) \leftarrow \mathsf{SimEnc}(1^\lambda, m)$. Then

$$H_\infty(C|\mathsf{CRS}) \geq \sum_{i \in [k']} H_\infty(C_i|U_i) \geq k'(\lambda)(e(\lambda) - \ell(\lambda))$$

where the first inequality follows from the fact that (C_i, U_i) are k' independent random variables, and the second inequality follows since $U_i \in \mathcal{L}_{\mathsf{pk}} - \hat{m}_i$ is supported over a set of size $2^{\ell(\lambda)}$. This shows that the SimEnc satisfies the entropy requirement.

Let $m = \{m_\lambda\}$ be any ensemble of messages of length $|m_\lambda| = k(\lambda)$. We show that the two distributions of (crs, c) are indistinguishable for $\mathsf{crs} \leftarrow \{0,1\}^{t(\lambda,k)}, c \leftarrow \mathsf{Enc}_{\mathsf{crs}}(1^\lambda, m)$ versus $(\mathsf{crs}, c) \leftarrow \mathsf{SimEnc}(1^\lambda, m)$. We do so via a sequence of hybrids.

Hybrid 0: This is the distribution of $\mathsf{crs} \leftarrow \{0,1\}^{t(\lambda,k)}, c \leftarrow \mathsf{Enc}_{\mathsf{crs}}(1^\lambda, m)$ where $m = (m_1, \ldots, m_{k'}) \in \{0,1\}^{r(\lambda) \cdot k'}$. The values are sampled as follows:
- For $i \in [k']$, choose $u_i \leftarrow \{0,1\}^{r'(\lambda)}$.
- Choose $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$.
- For $i \in [k']$, let $\hat{m}_i = \mathsf{embed}_{\mathsf{pk}}(m_i) \in \mathcal{R}_{\mathsf{pk}}$, $y_i := \mathsf{sam}(\mathsf{pk}; u_i) \in \mathcal{R}_{\mathsf{pk}}$ and $c_i \leftarrow \mathsf{Inv}_{\mathsf{td}}(y_i + \hat{m}_i)$.
- Output $(\mathsf{crs} = (u_1, \ldots, u_{k'}), c = (\mathsf{pk}, c_1, \ldots, c_{k'})$.

Hybrid 1: Instead of choosing $u_i \leftarrow \{0,1\}^{r'(\lambda)}$ and setting $y_i := \mathsf{sam}(\mathsf{pk}; u_i) \in \mathcal{R}_{\mathsf{pk}}$, we now choose $y_i \leftarrow \mathcal{R}_{\mathsf{pk}}$ and set $u_i \leftarrow \mathsf{explain}_{\mathsf{pk}}(y_i)$. That is, hybrid 1 is defined as follows:
- Choose $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$.
- For $i \in [k']$, choose $y_i \leftarrow \mathcal{R}_{\mathsf{pk}}$ and $u_i \leftarrow \mathsf{explain}_{\mathsf{pk}}(y_i)$.
- For $i \in [k']$, let $\hat{m}_i = \mathsf{embed}_{\mathsf{pk}}(m_i) \in \mathcal{R}_{\mathsf{pk}}$, $y_i := \mathsf{sam}(\mathsf{pk}; u_i) \in \mathcal{R}_{\mathsf{pk}}$ and $c_i \leftarrow \mathsf{Inv}_{\mathsf{td}}(y_i + \hat{m}_i)$.
- Output $(\mathsf{crs} = (u_1, \ldots, u_{k'}), c = (\mathsf{pk}, c_1, \ldots, c_{k'})$.

Hybrids 0 and 1 are statistically indistinguishable by the "oblivious sampling" property of the range $\mathcal{R}_{\mathsf{pk}}$ of the SLF*.

Hybrid 2: In hybrid 2, instead of choosing $y_i \leftarrow \mathcal{R}_{\mathsf{pk}}$ and setting $c_i \leftarrow \mathsf{Inv}_{\mathsf{td}}(y_i + \hat{m}_i)$, we choose $c_i \leftarrow D(\mathsf{pk})$ at random and set $y_i = f_{\mathsf{pk}}(c_i) - \hat{m}_i$. That is, hybrid 2 is defines as follows:
- Choose $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$.
- For $i \in [k']$, choose $c_i \leftarrow D(\mathsf{pk})$.
- For $i \in [k']$, let $\hat{m}_i = \mathsf{embed}_{\mathsf{pk}}(m_i) \in \mathcal{R}_{\mathsf{pk}}$, $y_i = f_{\mathsf{pk}}(c_i) - \hat{m}_i$ and $u_i \leftarrow \mathsf{explain}_{\mathsf{pk}}(y_i)$.
- Output $(\mathsf{crs} = (u_1, \ldots, u_{k'}), c = (\mathsf{pk}, c_1, \ldots, c_{k'})$.

Hybrids 1 and 2 are statistically indistinguishable by the requirement on the surjective mode of the SLF*, which ensures that the two distributions on (y_i, c_i) in hybrids 1 and 2 are indistinguishable.

Hybrid 3: In hybrid 3, instead of choosing $(\mathsf{pk}, \mathsf{td}) \leftarrow \mathsf{SurGen}(1^\lambda)$ we choose $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$. Note that td is never used hybrid 2. That is, hybrid 3 is defined as follows.

- Choose $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$.
- For $i \in [k']$, choose $c_i \leftarrow D(\mathsf{pk})$.
- For $i \in [k']$, let $\hat{m}_i = \mathsf{embed}_{\mathsf{pk}}(m_i) \in \mathcal{R}_{\mathsf{pk}}$, $y_i = f_{\mathsf{pk}}(c_i) - \hat{m}_i$ and $u_i \leftarrow \mathsf{explain}_{\mathsf{pk}}(y_i)$.
- Output $(\mathsf{crs} = (u_1, \dots, u_{k'}), c = (\mathsf{pk}, c_1, \dots, c_{k'})$.

Hybrids 2 and 3 are computationally indistinguishable by the indistinguishability requirement on the SLF.

Hybrid 4: In hybrid 4, if $f_{\mathsf{pk}}(c_i) \notin \mathcal{L}_{\mathsf{pk}}$ we set $y_i = L - \hat{m}_i$ where L is some arbitrary fixed element of $\mathcal{L}_{\mathsf{pk}}$. That is, hybrid 4 is defined as follows:

- Choose $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$.
- For $i \in [k']$, choose $c_i \leftarrow D(\mathsf{pk})$.
- For $i \in [k']$, let $\hat{m}_i = \mathsf{embed}_{\mathsf{pk}}(m_i) \in \mathcal{R}_{\mathsf{pk}}$. Let $y_i = f_{\mathsf{pk}}(c_i) - \hat{m}_i$ if $f_{\mathsf{pk}}(c_i) \in \mathcal{L}_{\mathsf{pk}}$ or else set $y_i = L - \hat{m}_i$ where L is some arbitrary element of $\mathcal{L}_{\mathsf{pk}}$. Let $u_i \leftarrow \mathsf{explain}_{\mathsf{pk}}(y_i)$.
- For $i \in [k']$, let $\hat{m}_i = \mathsf{embed}_{\mathsf{pk}}(m_i) \in \mathcal{R}_{\mathsf{pk}}$, $y_i = f_{\mathsf{pk}}(c_i) - \hat{m}_i$ and $u_i \leftarrow \mathsf{explain}_{\mathsf{pk}}(y_i)$.
- Output $(\mathsf{crs} = (u_1, \dots, u_{k'}), c = (\mathsf{pk}, c_1, \dots, c_{k'})$.

Hybrids 3,4 are indistinguishable by the lossy mode property of the SLF*. In particular, for a random $\mathsf{pk} \leftarrow \mathsf{LossyGen}(1^\lambda)$ and $c_i \leftarrow D(\mathsf{pk})$, the probability that $f_{\mathsf{pk}}(c_i) \notin \mathcal{L}_{\mathsf{pk}}$ is negligible.

Hybrid 4 is exactly the distribution of $(\mathsf{crs}, c) \leftarrow \mathsf{SimEnc}(1^\lambda, m)$. Therefore, we have shown that the two target distributions in Hybrid 0 and Hybrid 4 are indeed indistinguishable, which concludes the proof.

Corollary 1. *Under either the DCR or LWE assumptions, there exist good incompressible encodings that are locally decodable and achieve either:*

1. *Selective security in the CRS model, with a CRS of length $t(\lambda, k) = k(1 + o(1))$.*
2. *Composable security in the Random Oracle model.*

Furthermore the complexity of encoding/decoding is $k \cdot \mathrm{poly}(\lambda)$.

References

1. M. Ajtai. Generating hard instances of lattice problems (extended abstract). In *28th ACM STOC*, pages 99–108. ACM Press, May 1996.
2. J. Alwen, Y. Dodis, M. Naor, G. Segev, S. Walfish, and D. Wichs. Public-key encryption in the bounded-retrieval model. In H. Gilbert, editor, *EUROCRYPT 2010*, volume 6110 of *LNCS*, pages 113–134. Springer, Heidelberg, May / June 2010.
3. J. Alwen, Y. Dodis, and D. Wichs. Leakage-resilient public-key cryptography in the bounded-retrieval model. In Halevi [30], pages 36–54.

4. J. Alwen, S. Krenn, K. Pietrzak, and D. Wichs. Learning with rounding, revisited - new reduction, properties and applications. In R. Canetti and J. A. Garay, editors, *CRYPTO 2013, Part I*, volume 8042 of *LNCS*, pages 57–74. Springer, Heidelberg, Aug. 2013.
5. B. Applebaum, D. Cash, C. Peikert, and A. Sahai. Fast cryptographic primitives and circular-secure encryption based on hard learning problems. In Halevi [30], pages 595–618.
6. B. Auerbach, E. Kiltz, B. Poettering, and S. Schoenen. Lossy trapdoor permutations with improved lossiness. In M. Matsui, editor, *CT-RSA 2019*, volume 11405 of *LNCS*, pages 230–250. Springer, Heidelberg, Mar. 2019.
7. B. Barak, R. Shaltiel, and A. Wigderson. Computational analogues of entropy. In *Approximation, randomization, and combinatorial optimization*, volume 2764 of *Lecture Notes in Comput. Sci.*, pages 200–215. Springer, Berlin, 2003.
8. M. Bellare, D. Kane, and P. Rogaway. Big-key symmetric encryption: Resisting key exfiltration. In Robshaw and Katz [50], pages 373–402.
9. D. Boneh, J. Bonneau, B. Bünz, and B. Fisch. Verifiable delay functions. In H. Shacham and A. Boldyreva, editors, *CRYPTO 2018, Part I*, volume 10991 of *LNCS*, pages 757–788. Springer, Heidelberg, Aug. 2018.
10. Z. Brakerski, A. Langlois, C. Peikert, O. Regev, and D. Stehlé. Classical hardness of learning with errors. In D. Boneh, T. Roughgarden, and J. Feigenbaum, editors, *45th ACM STOC*, pages 575–584. ACM Press, June 2013.
11. Z. Brakerski, A. Lombardi, G. Segev, and V. Vaikuntanathan. Anonymous IBE, leakage resilience and circular security from new assumptions. In J. B. Nielsen and V. Rijmen, editors, *EUROCRYPT 2018, Part I*, volume 10820 of *LNCS*, pages 535–564. Springer, Heidelberg, Apr. / May 2018.
12. D. Cash, Y. Z. Ding, Y. Dodis, W. Lee, R. J. Lipton, and S. Walfish. Intrusion-resilient key exchange in the bounded retrieval model. In S. P. Vadhan, editor, *TCC 2007*, volume 4392 of *LNCS*, pages 479–498. Springer, Heidelberg, Feb. 2007.
13. E. Cecchetti, B. Fisch, I. Miers, and A. Juels. PIEs: Public incompressible encodings for decentralized storage. In L. Cavallaro, J. Kinder, X. Wang, and J. Katz, editors, *ACM CCS 2019*, pages 1351–1367. ACM Press, Nov. 2019.
14. Y. Chen, N. Genise, and P. Mukherjee. Approximate trapdoors for lattices and smaller hash-and-sign signatures. In S. D. Galbraith and S. Moriai, editors, *ASIACRYPT 2019, Part III*, volume 11923 of *LNCS*, pages 3–32. Springer, Heidelberg, Dec. 2019.
15. C. Cho, N. Döttling, S. Garg, D. Gupta, P. Miao, and A. Polychroniadou. Laconic oblivious transfer and its applications. In J. Katz and H. Shacham, editors, *CRYPTO 2017, Part II*, volume 10402 of *LNCS*, pages 33–65. Springer, Heidelberg, Aug. 2017.
16. J.-S. Coron, T. Holenstein, R. Künzler, J. Patarin, Y. Seurin, and S. Tessaro. How to build an ideal cipher: The indifferentiability of the Feistel construction. *Journal of Cryptology*, 29(1):61–114, Jan. 2016.
17. I. Damgård, C. Ganesh, and C. Orlandi. Proofs of replicated storage without timing assumptions. In A. Boldyreva and D. Micciancio, editors, *CRYPTO 2019, Part I*, volume 11692 of *LNCS*, pages 355–380. Springer, Heidelberg, Aug. 2019.
18. I. Damgård and M. Jurik. A generalisation, a simplification and some applications of Paillier's probabilistic public-key system. In K. Kim, editor, *PKC 2001*, volume 1992 of *LNCS*, pages 119–136. Springer, Heidelberg, Feb. 2001.
19. G. Di Crescenzo, R. J. Lipton, and S. Walfish. Perfectly secure password protocols in the bounded retrieval model. In Halevi and Rabin [31], pages 225–244.

20. Y. Dodis, A. Jain, T. Moran, and D. Wichs. Counterexamples to hardness amplification beyond negligible. In R. Cramer, editor, *TCC 2012*, volume 7194 of *LNCS*, pages 476–493. Springer, Heidelberg, Mar. 2012.
21. Y. Dodis, R. Ostrovsky, L. Reyzin, and A. D. Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. *SIAM J. Comput.*, 38(1):97–139, 2008.
22. Y. Dodis, S. P. Vadhan, and D. Wichs. Proofs of retrievability via hardness amplification. In O. Reingold, editor, *TCC 2009*, volume 5444 of *LNCS*, pages 109–127. Springer, Heidelberg, Mar. 2009.
23. S. Dziembowski. Intrusion-resilience via the bounded-storage model. In Halevi and Rabin [31], pages 207–224.
24. B. Fisch. Tight proofs of space and replication. Cryptology ePrint Archive, Report 2018/702, 2018. https://eprint.iacr.org/2018/702.
25. D. M. Freeman, O. Goldreich, E. Kiltz, A. Rosen, and G. Segev. More constructions of lossy and correlation-secure trapdoor functions. *Journal of Cryptology*, 26(1):39–74, Jan. 2013.
26. R. Garg, G. Lu, and B. Waters. New techniques in replica encodings with client setup. Cryptology ePrint Archive, Report 2020/617, 2020. https://eprint.iacr.org/2020/617.
27. C. Gentry, C. Peikert, and V. Vaikuntanathan. Trapdoors for hard lattices and new cryptographic constructions. In Ladner and Dwork [39], pages 197–206.
28. C. Gentry and D. Wichs. Separating succinct non-interactive arguments from all falsifiable assumptions. In L. Fortnow and S. P. Vadhan, editors, *43rd ACM STOC*, pages 99–108. ACM Press, June 2011.
29. S. Goldwasser, Y. T. Kalai, C. Peikert, and V. Vaikuntanathan. Robustness of the learning with errors assumption. In A. C.-C. Yao, editor, *ICS 2010*, pages 230–240. Tsinghua University Press, Jan. 2010.
30. S. Halevi, editor. *CRYPTO 2009*, volume 5677 of *LNCS*. Springer, Heidelberg, Aug. 2009.
31. S. Halevi and T. Rabin, editors. *TCC 2006*, volume 3876 of *LNCS*. Springer, Heidelberg, Mar. 2006.
32. J. Håstad, R. Impagliazzo, L. A. Levin, and M. Luby. A pseudorandom generator from any one-way function. *SIAM Journal on Computing*, 28(4):1364–1396, 1999.
33. C.-Y. Hsiao, C.-J. Lu, and L. Reyzin. Conditional computational entropy, or toward separating pseudoentropy from compressibility. In M. Naor, editor, *EUROCRYPT 2007*, volume 4515 of *LNCS*, pages 169–186. Springer, Heidelberg, May 2007.
34. A. Jain and K. Pietrzak. Parallel repetition for leakage resilience amplification revisited. In Y. Ishai, editor, *TCC 2011*, volume 6597 of *LNCS*, pages 58–69. Springer, Heidelberg, Mar. 2011.
35. A. Juels and B. S. Kaliski Jr. Pors: proofs of retrievability for large files. In P. Ning, S. De Capitani di Vimercati, and P. F. Syverson, editors, *ACM CCS 2007*, pages 584–597. ACM Press, Oct. 2007.
36. E. Kiltz, A. O'Neill, and A. Smith. Instantiability of RSA-OAEP under chosen-plaintext attack. In T. Rabin, editor, *CRYPTO 2010*, volume 6223 of *LNCS*, pages 295–313. Springer, Heidelberg, Aug. 2010.
37. P. Labs. Filecoin: A decentralized storage network. 2017.
38. P. Labs. Proof of replication. 2017.
39. R. E. Ladner and C. Dwork, editors. *40th ACM STOC*. ACM Press, May 2008.
40. A. B. Lewko and B. Waters. On the insecurity of parallel repetition for leakage resilience. In *51st FOCS*, pages 521–530. IEEE Computer Society Press, Oct. 2010.

41. U. M. Maurer, R. Renner, and C. Holenstein. Indifferentiability, impossibility results on reductions, and applications to the random oracle methodology. In M. Naor, editor, *TCC 2004*, volume 2951 of *LNCS*, pages 21–39. Springer, Heidelberg, Feb. 2004.
42. D. Micciancio and C. Peikert. Trapdoors for lattices: Simpler, tighter, faster, smaller. In D. Pointcheval and T. Johansson, editors, *EUROCRYPT 2012*, volume 7237 of *LNCS*, pages 700–718. Springer, Heidelberg, Apr. 2012.
43. M. Naor. On cryptographic assumptions and challenges (invited talk). In D. Boneh, editor, *CRYPTO 2003*, volume 2729 of *LNCS*, pages 96–109. Springer, Heidelberg, Aug. 2003.
44. P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In J. Stern, editor, *EUROCRYPT'99*, volume 1592 of *LNCS*, pages 223–238. Springer, Heidelberg, May 1999.
45. C. Peikert. Public-key cryptosystems from the worst-case shortest vector problem: extended abstract. In M. Mitzenmacher, editor, *41st ACM STOC*, pages 333–342. ACM Press, May / June 2009.
46. C. Peikert and A. Rosen. Efficient collision-resistant hashing from worst-case assumptions on cyclic lattices. In Halevi and Rabin [31], pages 145–166.
47. C. Peikert and B. Waters. Lossy trapdoor functions and their applications. In Ladner and Dwork [39], pages 187–196.
48. O. Regev. On lattices, learning with errors, random linear codes, and cryptography. In H. N. Gabow and R. Fagin, editors, *37th ACM STOC*, pages 84–93. ACM Press, May 2005.
49. T. Ristenpart, H. Shacham, and T. Shrimpton. Careful with composition: Limitations of the indifferentiability framework. In K. G. Paterson, editor, *EUROCRYPT 2011*, volume 6632 of *LNCS*, pages 487–506. Springer, Heidelberg, May 2011.
50. M. Robshaw and J. Katz, editors. *CRYPTO 2016, Part I*, volume 9814 of *LNCS*. Springer, Heidelberg, Aug. 2016.
51. H. Shacham and B. Waters. Compact proofs of retrievability. In J. Pieprzyk, editor, *ASIACRYPT 2008*, volume 5350 of *LNCS*, pages 90–107. Springer, Heidelberg, Dec. 2008.
52. M. van Dijk, A. Juels, A. Oprea, R. L. Rivest, E. Stefanov, and N. Triandopoulos. Hourglass schemes: how to prove that cloud files are encrypted. In T. Yu, G. Danezis, and V. D. Gligor, editors, *ACM CCS 2012*, pages 265–280. ACM Press, Oct. 2012.
53. D. Wichs. Barriers in cryptography with weak, correlated and leaky sources. In R. D. Kleinberg, editor, *ITCS 2013*, pages 111–126. ACM, Jan. 2013.
54. A. C.-C. Yao. Theory and applications of trapdoor functions (extended abstract). In *23rd FOCS*, pages 80–91. IEEE Computer Society Press, Nov. 1982.
55. M. Zhandry. The magic of ELFs. In Robshaw and Katz [50], pages 479–508.