Limits on the Power of Zero-Knowledge Proofs in Cryptographic Constructions

Zvika Brakerski[1], Jonathan Katz[2], Gil Segev[3], and Arkady Yerukhimovich[2]

[1] Weizmann Institute of Science, Rehovot, Israel
zvika.brakerski@weizmann.ac.il
[2] University of Maryland, College Park, MD, USA
{jkatz,arkady}@cs.umd.edu
[3] Microsoft Research, Mountain View, CA, USA
gil.segev@microsoft.com

Abstract. For over 20 years, black-box impossibility results have been used to argue the infeasibility of constructing certain cryptographic primitives (e.g., key agreement) from others (e.g., one-way functions). A widely recognized limitation of such impossibility results, however, is that they say nothing about the usefulness of (known) *non*black-box techniques. This is unsatisfying, as we would at least like to rule out constructions using the set of techniques we have at our disposal.

With this motivation in mind, we suggest a new framework for black-box constructions that encompasses constructions with a nonblack-box flavor: specifically, those that rely on *zero-knowledge proofs* relative to some oracle. We show that our framework is powerful enough to capture the Naor-Yung/Sahai paradigm for building a (shielding) CCA-secure public-key encryption scheme from a CPA-secure one, something ruled out by prior black-box separation results. On the other hand, we show that several black-box impossibility results still hold even in a setting that allows for zero-knowledge proofs.

1 Introduction

A central goal of theoretical cryptography is to explore relationships between various cryptographic primitives and, in particular, to show constructions of various "high-level" cryptographic objects (encryption schemes, key-agreement protocols, etc.) based on "low-level" cryptographic tools (such as one-way functions). This line of research has been very successful in many cases. In other cases, however, constructions of certain primitives from others are unknown: for example, we do not currently know how to construct public-key encryption schemes based on one-way functions. Given this failure, it is natural to wonder whether such constructions are inherently *impossible*. Unfortunately, we cannot rule out *all* such constructions as long as we believe that the object in question exists in the real world: if we believe that RSA encryption (say) is secure, then a valid construction of public-key encryption from any one-way function f consists of simply ignoring f and then outputting the code for the RSA encryption scheme. Yet this is clearly not what is intended.

In an effort to capture what is meant by a "natural" construction of one primitive from another, Impagliazzo and Rudich [15] formalized the notion of a *black-box* construction. Informally, a black-box construction of A from B is a construction of A that uses only the input/output characteristics of an implementation of B, but does not rely on any internal details as to how B is implemented. Moreover, A should be "secure" as long as B is "secure" (each in their respective senses). (We refer the reader to the work of Reingold, Trevisan, and Vadhan [19] for a more extensive definitional treatment.) Impagliazzo and Rudich show that there does not exist a black-box construction of key agreement from one-way functions, and since their work many more black-box impossibility results have been shown.

A recognized drawback of existing black-box impossibility results is that they say nothing regarding whether these results might be circumvented using *non*black-box techniques. While it is true that most constructions in cryptography are black-box, we have many examples of nonblack-box constructions as well. One striking example is given by the observation that all known constructions of CCA-secure public-key encryption schemes based on trapdoor permutations [18, 5, 20, 16] are, in fact, not black-box. (Interestingly, a partial black-box separation is known [11].) Other nonblack-box constructions include those of [6, 4, 3, 1, 9]; we refer the reader to [10] for further discussion and additional examples.

If black-box constructions are supposed to be representative of existing techniques, we should update our definition of what "black-box" means. In this paper, we propose a framework to do exactly this. Specifically, we suggest a model that incorporates a specific class of nonblack-box techniques: those that rely on zero-knowledge proofs. We accomplish this by augmenting the basic, black-box model — in which there is only an oracle $\mathcal{O}$ for some primitive — with a *zero-knowledge (ZK) oracle* that allows parties to prove statements in zero knowledge relative to $\mathcal{O}$. (Technically, a ZK oracle allows zero-knowledge proofs for any language in $\mathcal{NP}^{\mathcal{O}}$. We also find it simpler to work with a *witness-indistinguishability (WI) oracle*, but we show that a WI oracle implies zero-knowledge proofs in the settings we consider. In fact, although we do not define the notion of proofs of *knowledge*, our formulation can also be seen as providing that stronger property.) We call any construction using black-box access to $\mathcal{O}$ *and its associated WI oracle* an **augmented black-box** construction. Given primitives A and B, we can then ask whether there exists an *augmented* black-box construction of A from B; an impossibility result demonstrating that no such construction exists rules out a broader class of approaches to constructing one from the other. Of course, as with all impossibility results, such a result says nothing about whether some *other* nonblack-box techniques might apply (and, in fact, the nonblack-box results of, e.g., [3, 1] do not fall within our framework); nevertheless, impossibility results are still useful insofar as they show us where we must look if we hope to circumvent them.

Our contributions. We view the primary contribution of this paper as definitional and conceptual. In addition to putting forth the notion of augmented black-box constructions, however, we also show several technical results. To val-

idate our framework, we show that the Naor-Yung/Sahai [18, 20] (shielding) construction of CCA-secure public-key encryption from CPA-secure public-key encryption falls within our framework. (Such a construction is ruled out, in a black-box sense, by the result of Gertner et al. [11].) We note that several other existing nonblack-box constructions also fall within our framework, including those of [6, 4, 9]. This demonstrates that our framework meaningfully encompasses constructions that lie outside the standard black-box model.

On the negative side, we present two impossibility results for augmented black-box constructions. Generalizing the work of Impagliazzo and Rudich [15], we rule out augmented (fully) black-box constructions of key-agreement protocols with perfect completeness from one-way functions. (We leave the case of protocols without perfect completeness as an open problem.) Generalizing results of Haitner et al. [12, 13], we rule out augmented (fully) black-box constructions of statistically-hiding commitment schemes with low round complexity or low communication complexity from enhanced trapdoor permutations. Though it may seem "intuitively obvious" to the reader that zero-knowledge proofs cannot help in these settings, the challenge — as in all black-box impossibility proofs — is to prove this intuition. (In fact, under our initial modeling of a random WI proof system there *was* a construction of key agreement from one-way functions.)

Outline of the paper. In Section 2 we define the notion of augmented black-box constructions, and in Section 3 we show that our framework encompasses the Naor-Yung/Sahai paradigm for building CCA-secure public-key encryption from CPA-secure schemes. Our main technical results are in the sections that follow. We rule out augmented black-box constructions of key agreement from one-way functions in Section 4, and in Section 5 we prove lower bounds on the round complexity and communication complexity of augmented black-box constructions of statistically-hiding commitments from trapdoor permutations.

2 Augmented Black-Box Constructions

In this section we formally define our notion of *augmented black-box constructions*. Recall that our goal here is to model constructions that use an oracle $\mathcal{O}$ for some primitive as a black box, while also (possibly) using zero-knowledge proofs of $\mathcal{NP}$ statements relative to $\mathcal{O}$. To enable such proofs we introduce an additional set of oracles $(\mathcal{P}, \mathcal{V})$ implementing a "prover" and a "verifier", respectively. We find it easiest to model $(\mathcal{P}, \mathcal{V})$ as a *witness-indistinguishable* (WI) proof system [7], and to prove our impossibility results relative to oracles achieving this notion. In Section 2.2, however, we show that any WI proof system can be used to construct non-interactive zero-knowledge (NIZK) proofs in the common random string model, assuming the existence of one-way functions.

Fix an oracle $\mathcal{O} : \{0,1\}^* \to \{0,1\}^*$. For a language L, we say $L \in \mathcal{NP}^{\mathcal{O}}$ if there exists a polynomial-time oracle machine M running in time polynomial in its first input such that $x \in L$ if and only if there exists a witness w for which $M^{\mathcal{O}}(x, w)$ accepts. (We assume a valid witness w satisfies $|w| = |x|$ without loss

of generality.) For any $L \in \mathcal{NP}^{\mathcal{O}}$, we let R_L denote an $\mathcal{NP}$-relation associated with L, and we let $L_n \stackrel{\text{def}}{=} L \cap \{0,1\}^n$ and $R_n \stackrel{\text{def}}{=} \{(x,w) \mid (x,w) \in R_L \text{ and } x \in L_n\}$.

We now define what it means for a pair of oracles $(\mathcal{P}, \mathcal{V})$ to be a witness-indistinguishable proof system. (All adversaries are stateful by default.)

Definition 2.1. *Fix an oracle $\mathcal{O}$, a language $L \in \mathcal{NP}^{\mathcal{O}}$, and an $\mathcal{NP}$ relation R_L for L. An oracle $\mathcal{WI} = (\mathcal{P}, \mathcal{V})$ is a* proof system *for R_L if the following hold:*

- **Perfect completeness:** *For any $n \in \mathbb{N}$, $(x,w) \in R_n$, and $r \in \{0,1\}^n$, it holds that $\mathcal{V}_n(x, \mathcal{P}_n(x,w,r)) = 1$.*
- **Perfect soundness:** *For any $x \notin L$ and any π, it holds that $\mathcal{V}_n(x,\pi) = 0$.*

$\mathcal{WI}$ is witness-indistinguishable (WI) *if additionally:*

- **Witness indistinguishability:** *For every probabilistic polynomial-time $\mathcal{A}$, it holds that $|\Pr[\mathsf{ExptWI}_{\mathcal{A}}(n) = 1] - 1/2|$ is negligible, where $\mathsf{ExptWI}_{\mathcal{A}}(n)$ is defined as follows:*

$$\begin{array}{rcl} (x, w_0, w_1) \leftarrow \mathcal{A}^{\mathcal{O},\mathcal{WI}}(1^n); b \leftarrow \{0,1\}; & & \text{if } (x,w_0),(x,w_1) \in R_n \\ r \leftarrow \{0,1\}^n; \pi \leftarrow \mathcal{P}_n(x, w_b, r) & : & \text{output 1 iff } b' = b \\ b' = \mathcal{A}^{\mathcal{O},\mathcal{WI}}(1^n, \pi) & & \text{else, output a random bit} \end{array}$$

When the relation R_L is irrelevant for the discussion at hand, or is clear from the context, we may abuse terminology and call $\mathcal{WI}$ a WI proof system for L. We say that $\mathcal{WI}$ is a WI proof system for $\mathcal{NP}^{\mathcal{O}}$ if it is a WI proof system for the $\mathcal{NP}^{\mathcal{O}}$-complete language CIRCUIT-SAT$^{\mathcal{O}}$ (the set of satisfiable circuits with $\mathcal{O}$ gates) under the natural relation R_L.

We now define our notion of black-box reductions using a base oracle $\mathcal{O}$ and a WI oracle $\mathcal{WI}$ for $\mathcal{NP}^{\mathcal{O}}$. The definitions and terminology are adapted from [19].

Definition 2.2 (Augmented fully black-box construction). *There is an* augmented fully black-box construction *of primitive Q from primitive P if there exist probabilistic polynomial-time oracle machines G and S such that:*

- *For any $\mathcal{O}, \mathcal{WI}$ such that $\mathcal{O}$ implements P, and $\mathcal{WI}$ is a proof system for $\mathcal{NP}^{\mathcal{O}}$, the algorithm $G^{\mathcal{O},\mathcal{WI}}$ implements Q.*
- *For any $\mathcal{O}, \mathcal{WI}$ and any (possibly inefficient) adversary $\mathcal{A}^{\mathcal{O},\mathcal{WI}}$ that breaks the Q-security of $G^{\mathcal{O},\mathcal{WI}}$, the adversary $S^{\mathcal{A},\mathcal{O},\mathcal{WI}}$ breaks the P-security of $\mathcal{O}$ or the witness indistinguishability of $\mathcal{WI}$.*

Definition 2.3 (Augmented semi-black-box construction). *There is an* augmented semi-black-box construction *of primitive Q from primitive P if there exists a probabilistic polynomial-time oracle machine G such that:*

- *For any $\mathcal{O}, \mathcal{WI}$ such that $\mathcal{O}$ implements P, and $\mathcal{WI}$ is a proof system for $\mathcal{NP}^{\mathcal{O}}$, the algorithm $G^{\mathcal{O},\mathcal{WI}}$ implements Q.*
- *For any $\mathcal{O}, \mathcal{WI}$ and any probabilistic polynomial-time adversary $\mathcal{A}^{\mathcal{O},\mathcal{WI}}$ that breaks the Q-security of $G^{\mathcal{O},\mathcal{WI}}$, there is a probabilistic polynomial-time S such that $S^{\mathcal{O},\mathcal{WI}}$ breaks the P-security of $\mathcal{O}$ or the witness indistinguishability of $\mathcal{WI}$.*

We remark that our notions of augmented black-box constructions are not transitive: i.e., if there is an augmented black-box construction of Q from P, and an augmented black-box construction of R from Q, this does not imply that there is an augmented black-box construction of R from P. (On the other hand, if either of the given constructions is black-box, that does imply an augmented black-box construction of R from P.) The reason is that $\mathcal{WI}$ enables proofs for $\mathcal{NP}^{\mathcal{O}}$ but not $\mathcal{NP}^{\mathcal{O},\mathcal{WI}}$. While it is true that Definition 2.1 can be meaningfully changed to allow for proofs of $\mathcal{NP}^{\mathcal{O},\mathcal{WI}}$, doing so introduces technical issues (due to circularity) and we were unable to prove our separation results with respect to such a definition. We leave this as an interesting open question.

2.1 Instantiating a WI Proof System

For arbitrary $\mathcal{O}$, we show how to instantiate a WI proof system for $\mathcal{NP}^{\mathcal{O}}$. We begin by describing a distribution such that an oracle sampled according to this distribution is a WI proof system for $\mathcal{NP}^{\mathcal{O}}$ with overwhelming probability (Lemma 2.5). We then show that this implies that measure 1 of the oracles under this distribution constitute a WI proof system for $\mathcal{NP}^{\mathcal{O}}$ (Lemma 2.6). Throughout this section, we take L to be CIRCUIT-SAT$^{\mathcal{O}}$.

It is convenient to view the (infinite) oracle $\mathcal{WI}$ as a sequence of oracles $\{\mathcal{WI}_n = (\mathcal{P}_n, \mathcal{V}_n)\}_{n\in\mathbb{N}}$, one for each input length. Consider the distribution over $\mathcal{WI}$ where, for each n, the distribution over $\mathcal{WI}_n$ is defined as follows:

Prover oracle: $\mathcal{P}_n$ is a random function $\mathcal{P}_n : \{0,1\}^{3n} \to \{0,1\}^{7n}$ whose inputs are parsed as tuples of the form $(x, w, r) \in \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^n$. Note that $\mathcal{P}_n$ is defined for all such tuples (x, w, r) of the appropriate length, and not only for those satisfying $(x, w) \in R_L$ (i.e., $\mathcal{P}_n$ does not check whether $(x, w) \in R_L$).

Verifier oracle: The verifier oracle is a function $\mathcal{V}_n : \{0,1\}^{8n} \to \{0,1\}$, whose inputs are parsed as pairs of the form $(x, \pi) \in \{0,1\}^n \times \{0,1\}^{7n}$. The function is defined as:

$$\mathcal{V}_n(x,\pi) = \begin{cases} 1 \text{ if } \exists(w,r) \text{ s.t. } \pi = \mathcal{P}_n(x,w,r) \ \wedge \ (x,w) \in R_L \\ 0 \text{ otherwise} \end{cases}$$

Note that $\mathcal{WI}$ sampled as above is always a proof system. It remains to show that witness indistinguishability holds with overwhelming probability. We begin by proving that, for oracles distributed as above, it is essentially impossible to "spoof" a proof. That is, for n large enough, the only way to generate a proof π such that $\mathcal{V}_n(x, \pi) = 1$ is by querying $\mathcal{P}_n$. This property of the $\mathcal{WI}$ oracle will also be useful later.

Lemma 2.4. *For an oracle algorithm $\mathcal{A}^{\mathcal{O},\mathcal{WI}}$, let Spoof_n be the event that $\mathcal{A}$ makes a query $\mathcal{V}_n(x, \pi)$ that returns 1, yet π was not output by a previous query $\mathcal{P}_n(x, w, \star)$ with $(x, w) \in R_L$. For any $\mathcal{O}$, any $\mathcal{A}$ making at most q oracle queries, and any n, the probability of Spoof_n is at most $q \cdot 2^{-4n}$ (where the probability is taken over choice of $\mathcal{WI}$ according to the distribution above).*

Proof (sketch). We drop the subscript n for ease of presentation. There are at most 2^{3n} elements in the range of $\mathcal{P}$ and these are distributed uniformly in a space of size 2^{7n}. Since $\mathcal{P}$ is chosen independently of $\mathcal{O}$, queries to $\mathcal{O}$ give no information about the range of $\mathcal{P}$. Each $\mathcal{P}$-query reveals one point in the range of $\mathcal{P}$, but as the other points in the range are chosen independently this does not help to find another element in the range. The probability that any particular query $\mathcal{V}(x,\pi)$ returns 1 if π was not previously output by $\mathcal{P}$ is at most 2^{-4n}. Taking a union bound over all the queries of $\mathcal{A}$ gives the desired result.

Lemma 2.5. *For any oracle $\mathcal{O}$, every probabilistic polynomial-time oracle machine $\mathcal{A}$, and n large enough:*

$$\left|\Pr\left[\mathsf{ExptWI}_{\mathcal{A}}(n)=1\right]-\frac{1}{2}\right|\leq 2^{-n/2},$$

where $\mathsf{ExptWI}_{\mathcal{A}}(n)$ is as in Definition 2.1, and the above probability is also taken over the choice of $\mathcal{WI}$.

Proof. Consider some value of n, and fix the values of $\mathcal{WI}$ other than $\mathcal{WI}_n$. Assume without loss of generality that $\mathcal{A}(1^n)$ outputs values (x,w_0,w_1) with $(x,w_0),(x,w_1)\in R_n$. Then $\mathcal{A}$ is given a proof π and has to identify whether w_0 or w_1 was used to generate it. We observe that for all $k\neq n$ the output of any query to $\mathcal{P}_k$ or $\mathcal{V}_k$ is independent of the bit b. Therefore from this point on we focus on queries to $\mathcal{P}_n$ and $\mathcal{V}_n$. Let q be the total number of oracle queries made by $\mathcal{A}$. We may assume that $\mathcal{A}$ does not query $\mathcal{V}_n$ since it can simulate this oracle by itself to within statistical difference at most 2^{-n} (for n large enough). Indeed, there are three types of queries to $\mathcal{V}_n$:

- The query $\mathcal{V}_n(x,\pi)$. In this case, the output is 1.
- Queries of the form $\mathcal{V}_n(x,\pi')$, where π' was output by a previous query $\mathcal{P}_n(x,w,\star)$ with $(x,w)\in R_n$. Once again, in this case the output is 1. Note that $\mathcal{A}$ can check in polynomial time whether $(x,w)\in R_n$.
- All other queries to $\mathcal{V}_n$. In this case, Lemma 2.4 shows that the output of all these queries is 0 except with probability at most $q\cdot 2^{-4n}$, which is bounded by 2^{-n} for n sufficiently large.

Given the above, and the fact that $\mathcal{P}_n$ is chosen at random, it follows that $\mathcal{A}$ cannot distinguish which witness was used with probability better than $q\cdot 2^{-n}$, which is bounded by $2^{-n/2}$ for n sufficiently large. The lemma follows.

Lemma 2.6. *Fix an oracle $\mathcal{O}$. For measure 1 of the oracles $\mathcal{WI}$ under the distribution defined above, $\mathcal{WI}$ is a witness-indistinguishable proof system for L.*

Proof. Completeness and soundness always hold, and so we must only prove witness indistinguishability. To do so we apply a standard argument using the Borel-Cantelli lemma for reversing the order of quantifiers in Lemma 2.5.

Fix $\mathcal{O}$. For any $n \in \mathbb{N}$ and any probabilistic polynomial-time $\mathcal{A}$, denote by $E_{n,\mathcal{A}}$ the event in which $\mathcal{WI}$ is chosen such that

$$\left| \Pr\left[\mathsf{ExptWI}_{\mathcal{A}}(n) = 1\right] - \frac{1}{2} \right| > 2^{-n/3}.$$

Lemma 2.5 and an averaging argument imply that for any $\mathcal{A}$ and sufficiently large n the probability of $E_{n,\mathcal{A}}$ is at most $1/n^2$. Then $\sum_n \Pr[E_{n,\mathcal{A}}]$ is finite, and so the Borel-Cantelli lemma implies that the probability over choice of $\mathcal{WI}$ that event $E_{n,\mathcal{A}}$ occurs for infinitely many values of n is zero. Thus, for large enough n and measure 1 of the oracles under the distribution in question we have

$$\left| \Pr\left[\mathsf{ExptWI}_{\mathcal{A}}(n) = 1\right] - \frac{1}{2} \right| \leq 2^{-n/3}.$$

This holds for any specific $\mathcal{A}$, and therefore by removing a set of measure 0 for each of the (countably many) machines $\mathcal{A}$ we obtain that for measure 1 of the oracles $\mathcal{WI}$ it holds that for *all* probabilistic polynomial-time $\mathcal{A}$ the quantity $\left|\Pr\left[\mathsf{ExptWI}_{\mathcal{A}}(n) = 1\right] - \frac{1}{2}\right|$ is negligible.

Before concluding this section we prove a technical result regarding oracles $\mathcal{WI}$ sampled according to the distribution described earlier. We show that if f is one-way relative to $\mathcal{O}$, then for measure 1 of the oracles $\mathcal{WI}$ under the distribution defined above f remains one-way relative to $(\mathcal{O}, \mathcal{WI})$.

Lemma 2.7. *Let f be a polynomial-time oracle machine such that $f^{\mathcal{O}}$ is one-way relative to $\mathcal{O}$. Then for measure 1 of the oracles $\mathcal{WI}$ under the distribution defined above, $f^{\mathcal{O}}$ is one-way relative to $(\mathcal{O}, \mathcal{WI})$.*

Proof. It suffices to show that for any PPT $\mathcal{A}$ the probability that $\mathcal{A}^{\mathcal{O},\mathcal{WI}}$ inverts $f^{\mathcal{O}}$ is negligible, where the probability is also taken over choice of $\mathcal{WI}$. We can then proceed as in Lemma 2.6 to obtain the stated result.

Assume toward a contradiction that there exists an algorithm $\mathcal{A}$ and a polynomial $p(n) \geq n$ such that the running time of $\mathcal{A}$ is bounded by $p(n)$ and, for infinitely many n, it holds that $A^{\mathcal{O},\mathcal{WI}}$ inverts $f^{\mathcal{O}}$ with probability at least $1/p(n)$ when $\mathcal{WI}$ is chosen at random. We show how to construct a PPT algorithm $\hat{\mathcal{A}}$ such that $\hat{\mathcal{A}}^{\mathcal{O}}$ inverts $f^{\mathcal{O}}$ with inverse-polynomial probability for infinitely many values of n, a contradiction.

$\hat{\mathcal{A}}(1^n, y)$ runs $\mathcal{A}(1^n, y)$, simulating the $\mathcal{WI}$ oracle for $\mathcal{A}$ as follows. Let $k^* = \log p(n)$. Algorithm $\hat{\mathcal{A}}$ samples $\mathcal{WI}_k = (\mathcal{P}_k, \mathcal{V}_k)$ according to the prescribed distribution for all $k \leq k^*$, and these are used to (perfectly) simulate $\{\mathcal{WI}_k\}_{k \leq k^*}$ to $\mathcal{A}$. Thus, we now only need to deal with the queries of $\mathcal{A}$ to $\mathcal{WI}_k$ for $k > k^*$. When $\mathcal{A}$ queries $\mathcal{P}_k(x, w, r)$, then $\hat{\mathcal{A}}$ returns a random $\pi \in \{0,1\}^{7k}$ as the result. When $\mathcal{A}$ queries $\mathcal{V}_k(x, \pi)$ then $\hat{\mathcal{A}}$ first checks to see whether there was any prior query $\mathcal{P}_k(x, w, \star) = \pi$ with $(x, w) \in R_L$. If not, then $\hat{\mathcal{A}}$ returns 0 in response to this $\mathcal{V}_k$-query. Otherwise, $\hat{\mathcal{A}}$ returns 1.

It follows from Lemma 2.4 that $\hat{\mathcal{A}}$'s simulation of $\mathcal{A}$ degrades the latter's probability of inversion by at most $1/2p(n)$. This implies that $\hat{\mathcal{A}}^{\mathcal{O}}$ inverts $f^{\mathcal{O}}$ with probability at least $1/2p(n)$ for infinitely many values of n, a contradiction.

2.2 Zero-Knowledge Proofs

We define a notion of zero knowledge, and then discuss appropriate conditions under which zero-knowledge (ZK) proofs can be constructed from WI proofs. In our context, zero knowledge is most easily expressed in terms of non-interactive zero knowledge in the common random string model.

Definition 2.8. *Fix an oracle $\mathcal{O}$ and a language $L \in \mathcal{NP}^{\mathcal{O}}$. An oracle $\mathcal{ZK} = (\mathcal{P}, \mathcal{V})$ is a* proof system in the common random string model *for L with relation R_L if there is a polynomial ℓ such that the following hold:*

- **Perfect completeness:** *For all $n \in \mathbb{N}$, all $(x, w) \in R_n$, all $\mathsf{crs} \in \{0,1\}^{\ell(n)}$, and all $r \in \{0,1\}^n$, we have $\mathcal{V}(\mathsf{crs}, x, \mathcal{P}(\mathsf{crs}, x, w, r)) = 1$.*
- **Statistical soundness:** *With all but negligible probability over choice of $\mathsf{crs} \in \{0,1\}^{\ell(n)}$, there do not exist $x \notin L_n$ and π such that $\mathcal{V}(\mathsf{crs}, x, \pi) = 1$.*

$\mathcal{ZK}$ is a non-interactive zero-knowledge (NIZK) proof system *if additionally:*

- **Black-box (adaptive) zero knowledge:** *There exists a* PPT simulator *$\mathcal{S} \stackrel{\text{def}}{=} (\mathcal{S}_1, \mathcal{S}_2)$ such that for all probabilistic polynomial-time $\mathcal{A}$ the following is negligible:*

$$\left| \Pr\left[\begin{array}{c} \mathsf{crs} \leftarrow \{0,1\}^{\ell(n)}; \\ (x,w) \leftarrow \mathcal{A}^{\mathcal{O},\mathcal{ZK}}(\mathsf{crs}); \\ r \leftarrow \{0,1\}^n; \\ \pi \leftarrow \mathcal{P}(\mathsf{crs}, x, w, r) \end{array} : \mathcal{A}^{\mathcal{O},\mathcal{ZK}}(\pi) = 1 \wedge (x,w) \in R_n \right] \right.$$
$$\left. - \Pr\left[\begin{array}{l} (\mathsf{crs}, s) \leftarrow \mathcal{S}_1^{\mathcal{O},\mathcal{ZK}}(1^n); \\ (x,w) \leftarrow \mathcal{A}^{\mathcal{O},\mathcal{ZK}}(\mathsf{crs}); : \mathcal{A}^{\mathcal{O},\mathcal{ZK}}(\pi') = 1 \wedge (x,w) \in R_n \\ \pi' \leftarrow \mathcal{S}_2^{\mathcal{A},\mathcal{O},\mathcal{ZK}}(s, x) \end{array} \right] \right| .$$

Constructing NIZK proofs from WI proofs. Fix an oracle $\mathcal{O}$, and let $\mathcal{WI} = (\mathcal{P}, \mathcal{V})$ be a WI proof system for $L = \text{CIRCUIT-SAT}^{\mathcal{O}}$. We show that if a one-way function $f^{\mathcal{O}}$ exists relative to $\mathcal{O}, \mathcal{WI}$, then we can construct an NIZK proof system for $\mathcal{NP}^{\mathcal{O}}$.

Assume $f^{\mathcal{O}}$ is one-way relative to $\mathcal{O}, \mathcal{WI}$. Using f, we can construct, in a black-box fashion, a pseudorandom generator $G^{\mathcal{O}} : \{0,1\}^n \to \{0,1\}^{2n}$ (see [14]). Define the following language $L' \in \mathcal{NP}^{\mathcal{O}}$:

$$L' \stackrel{\text{def}}{=} \left\{ (x, \mathsf{crs}) \text{ s.t. } \exists w \in \{0,1\}^n \text{ for which } (x,w) \in R_L \text{ or } \mathsf{crs} = G^{\mathcal{O}}(w) \right\}.$$

A zero-knowledge proof that $x \in L$ can then be constructed [7] by giving a witness-indistinguishable proof that $(x, \mathsf{crs}) \in L'$. In more detail, given a WI proof system $(\mathcal{P}, \mathcal{V})$ for L, consider the following proof system $(\mathcal{P}_{\mathcal{ZK}}, \mathcal{V}_{\mathcal{ZK}})$ for L:

Prover $\mathcal{P}_{\mathcal{ZK}}$: Given crs, x, w with $\mathsf{crs} \in \{0,1\}^{2n}$ and $(x,w) \in R_n$, set $x' = (x, \mathsf{crs})$ and note that $(x', w) \in L'$. Use a Levin reduction to the $\mathcal{NP}^{\mathcal{O}}$-complete language L to obtain $(\hat{x}, \hat{w}) \in L$. Choose $r \leftarrow \{0,1\}^{|\hat{x}|}$ and return the proof $\pi = \mathcal{P}(\hat{x}, \hat{w}, r)$.

Verifier $\mathcal{V}_{\mathcal{ZK}}$: Given crs, x, π, set $x' = (x, \mathsf{crs})$ and use a Levin reduction to the $\mathcal{NP}^{\mathcal{O}}$-complete language L to obtain $\hat{x}$. Then output $\mathcal{V}(\hat{x}, \pi)$.

Theorem 2.9. *If $(\mathcal{P}, \mathcal{V})$ is a WI proof system for L, then $(\mathcal{P}_{\mathcal{ZK}}, \mathcal{V}_{\mathcal{ZK}})$ is an NIZK proof system for L.*

Proof. Completeness is immediate, and statistical soundness of $(\mathcal{P}_{\mathcal{ZK}}, \mathcal{V}_{\mathcal{ZK}})$ follows from the soundness of $(\mathcal{P}, \mathcal{V})$ and the fact that a uniform $\mathsf{crs} \in \{0,1\}^{2n}$ is in the range of G with only negligible probability.

A simulator $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2)$ is given as follows. $\mathcal{S}_1(1^n)$ chooses $w \leftarrow \{0,1\}^n$ computes $\mathsf{crs} = G^{\mathcal{O}}(w)$, and then outputs (crs, w). Given x, simulator $\mathcal{S}_2$ sets $x' = (x, \mathsf{crs})$, applies a Levin reduction to (x', w) to obtain $(\hat{x}, \hat{w}) \in L$, chooses $r \leftarrow \{0,1\}^{|\hat{x}|}$, and outputs $\pi = \mathcal{P}(\hat{x}, \hat{w}, r)$.

The fact that $\mathcal{S}$ provides a good simulation follows from pseudorandomness of G relative to $\mathcal{O}, \mathcal{WI}$, and witness indistinguishability of $\mathcal{WI}$.

3 An Augmented Black-Box Construction

Here we show that the Naor-Yung/Sahai construction of CCA-secure public-key encryption from CPA-secure public-key encryption can be cast as an augmented fully black-box construction. This result is not surprising; the point is to demonstrate that our framework does, indeed, capture constructions that go beyond the usual black-box ones. In particular, the construction is *shielding* in the terminology of [11], something ruled out in that same work in a black-box sense.

Let $\mathcal{O} = (G, E, D)$ be a public-key encryption scheme (with perfect correctness), and let $\mathcal{WI} = (\mathcal{P}, \mathcal{V})$ be a WI proof system for $\mathcal{NP}^{\mathcal{O}}$. Assume $\mathcal{O}$ is CPA-secure relative to $\mathcal{O}, \mathcal{WI}$. As noted in Section 2.2, we can use $\mathcal{WI}$ to construct an NIZK proof system $(\mathcal{P}_{\mathcal{ZK}}, \mathcal{V}_{\mathcal{ZK}})$ for $\mathcal{NP}^{\mathcal{O}}$. (Existence of CPA-secure encryption implies existence of a one-way function.) Moreover, we can use the results of Sahai [20] to transform $(\mathcal{P}_{\mathcal{ZK}}, \mathcal{V}_{\mathcal{ZK}})$ into a *simulation-sound* NIZK proof system $\mathsf{ss}\mathcal{ZK} = (\mathcal{P}_{\mathsf{ss}\mathcal{ZK}}, \mathcal{V}_{\mathsf{ss}\mathcal{ZK}})$ for $\mathcal{NP}^{\mathcal{O}}$. (We remark that for $\mathcal{WI}$ sampled according to the distribution described in Section 2.1, the NIZK proof system $(\mathcal{P}_{\mathcal{ZK}}, \mathcal{V}_{\mathcal{ZK}})$ would already satisfy simulation soundness with overwhelming probability. However, here we want a construction starting from *any* WI proof system.) For notational convenience, we will treat $\mathsf{ss}\mathcal{ZK}$ as an NIZK proof system for the specific language

$$L \stackrel{\text{def}}{=} \{(c_1, c_2, pk_1, pk_2) \mid \exists m, r_1, r_2 : c_1 = E^{\mathcal{O}}_{pk_1}(m; r_1) \bigwedge c_2 = E^{\mathcal{O}}_{pk_2}(m; r_2)\}.$$

We now describe the construction of a CCA-secure encryption scheme:

KeyGen $\mathcal{G}^{\mathcal{O}, \mathsf{ss}\mathcal{ZK}}$: Compute $(pk_1, sk_1) \leftarrow G(1^n)$ and $(pk_2, sk_2) \leftarrow G(1^n)$. Then choose $\mathsf{crs} \leftarrow \{0,1\}^{\ell(n)}$ and set $PK = (pk_1, pk_2, \mathsf{crs})$ and $SK = (sk_1, sk_2)$.

Encryption $\mathcal{E}^{\mathcal{O}, \mathsf{ss}\mathcal{ZK}}$: To encrypt plaintext m, choose $r_1, r_2, r \leftarrow \{0,1\}^n$ and then compute the ciphertexts $c_1 = E_{pk_1}(m; r_1)$ and $c_2 = E_{pk_2}(m; r_2)$. Set $x = (c_1, c_2, pk_1, pk_2)$ and $w = (m, r_1, r_2)$ and generate an NIZK proof $\pi = \mathcal{P}_{\mathsf{ss}\mathcal{ZK}}(\mathsf{crs}, x, w, r)$. Output (c_1, c_2, π).

Decryption $\mathcal{D}^{\mathcal{O}, \mathsf{ss}\mathcal{ZK}}$: To decrypt (c_1, c_2, π), set $x = (c_1, c_2, pk_1, pk_2)$ and check that $\mathcal{V}_{\mathsf{ss}\mathcal{ZK}}(\mathsf{crs}, x, \pi) = 1$. If not, output $\perp$. Otherwise, output $m = D_{sk_1}(c_1)$.

The following theorem follows from [20, Theorem 4.1].

Theorem 3.1. *Let $\mathcal{O}$ be an encryption scheme (with perfect correctness) that is CPA-secure relative to $\mathcal{O}, \mathcal{WI}$. Then the above is an augmented fully black-box construction of a CCA-secure encryption scheme from $\mathcal{O}$.*

4 An Impossibility Result for Key Agreement

In this section, we rule out augmented black-box constructions of key agreement with perfect completeness from one-way functions. (We conjecture that the result extends to the case of imperfect completeness, but we were unable to prove this.) For the remainder of this section, we only consider 1-bit key-agreement protocols with perfect completeness.

Say (A, B) is a pair of polynomial-time oracle algorithms that is an augmented black-box construction of key agreement from one-way functions. Then:

- For any $\mathcal{O}, \mathcal{WI}$ such that $\mathcal{WI}$ is a proof system for $\mathcal{NP}^{\mathcal{O}}$ and all n, following an execution between $A^{\mathcal{O},\mathcal{WI}}(1^n)$ and $B^{\mathcal{O},\mathcal{WI}}(1^n)$ both parties agree on a common bit $k \in \{0,1\}$.
- Given (A, B) and E, define the advantage of E by the following experiment:
  1. $A^{\mathcal{O},\mathcal{WI}}(1^n)$ and $B^{\mathcal{O},\mathcal{WI}}(1^n)$ interact, resulting in a shared key k and a transcript T.
  2. E is given T, and outputs a bit k'.

  The advantage of E is $|\Pr[k' = k] - 1/2|$.

  For any $\mathcal{O}$ and $\mathcal{WI}$ such that $\mathcal{O}$ is one-way relative to $(\mathcal{O}, \mathcal{WI})$ and $\mathcal{WI}$ is a WI proof system for $\mathcal{NP}^{\mathcal{O}}$, every unbounded algorithm E making at most polynomially many queries to $\mathcal{O}$ and $\mathcal{WI}$ has negligible advantage.

To prove that no augmented (fully) black-box construction of key agreement from one-way functions exists, fix some (A, B) and consider an execution of (A, B) when $\mathcal{O}$ is chosen at random and $\mathcal{WI}$ is chosen as described in Section 2.1. A random oracle is one-way [15], and Lemma 2.7 shows that it remains one-way in the presence of $\mathcal{WI}$ chosen from the specified distribution. Moreover, by Lemma 2.6 we have that $\mathcal{WI}$ is a WI proof system for $\mathcal{NP}^{\mathcal{O}}$. We note that even though these lemmas are stated with respect to polynomial time adversaries, since our proofs relativize, they also hold for computationally unbounded adversaries making at most polynomially many oracle queries. Thus, if (A, B) were an augmented black-box construction of key-agreement from one-way functions, then for any unbounded algorithm E making at most polynomially many oracle queries that has non-negligible advantage, there should exist a polynomial time machine $S^{E,\mathcal{O},\mathcal{WI}}$ that inverts $\mathcal{O}$ or breaks the witness indistinguishability of $\mathcal{WI}$. However, since S makes at most polynomially many queries to $\mathcal{O}, \mathcal{WI}$ (even indirectly through E), such an S does not exist. Therefore, every unbounded algorithm E making at most polynomially many queries to $\mathcal{O}$ and $\mathcal{WI}$ should have negligible advantage. However, we show an explicit E for which this is not the case, thus proving that no augmented (fully) black-box construction

of key agreement from one-way functions exists. (In fact, our attack works for *any* oracles $\mathcal{O}, \mathcal{WI}$, not just those chosen according to the distributions stated.) E can be made efficient if $\mathcal{P} = \mathcal{NP}$; thus any augmented semi-black-box construction of key agreement from one-way functions would imply $\mathcal{P} \neq \mathcal{NP}$.

4.1 Breaking Key Agreement Relative to a Random Oracle

In this section we provide a warmup for our main proof by ruling out (standard) black-box constructions of key agreement from one-way functions. This proof may also be of independent interest for pedagogical purposes as a simplified version of the proofs in [15, 2]. Note, however, that we prove a weaker result: we only rule out constructions of key-agreement protocols with perfect completeness based on one-way functions.

Let (A, B) be a construction of key agreement from one-way functions. Let q_A (resp., q_B) be a polynomial upper bound on the number of queries made by A (resp., B). Consider an attacker E defined as follows. E, given a transcript T of an execution of (A, B) in the presence of a random oracle $\mathcal{O}$, maintains a set $Q(E)$ of query/answer pairs for $\mathcal{O}$, and a multiset of candidate keys K, both initialized to $\emptyset$. Then E runs $2q_B + 1$ iterations of the following attack:

- *Simulation phase:* E finds a view of A consistent with the given transcript and with $Q(E)$. This view contains the randomness r_A used by A, as well as a set of oracle queries/answers $\hat{Q}(A)$ made by A. The set $\hat{Q}(A)$ is chosen to be consistent with any queries/answers in $Q(E)$, but it need not be consistent with the true oracle $\mathcal{O}$.
  Let k denote the key computed by A in the view. Then E adds k to K.
- *Update phase:* E makes all queries in $\hat{Q}(A) \setminus Q(E)$ to the true oracle $\mathcal{O}$, and adds the resulting query/answer pairs to $Q(E)$.

Following the above, E has a multiset K of $2q_B + 1$ possible keys. E outputs the majority value in K.

In each iteration E makes at most q_A queries to $\mathcal{O}$. Thus, E makes $O(q_A \cdot q_B)$ queries overall. We claim that E outputs the key computed by A and B with probability 1. Toward this, we first prove the following:

Claim 4.1. *Let k denote the actual key computed by A and B in an execution of the protocol. Then in each iteration of the attack, either E adds k to K, or E adds to $Q(E)$ one of the queries made by B in the real execution.*

Proof. Let $Q(B)$ denote the queries made by B in the real execution of the protocol. In a given iteration, there are two possibilities. If $\hat{Q}(A) \cap Q(B) \not\subseteq Q(E)$, then we are done since E makes all queries in $\hat{Q}(A) \setminus Q(E)$ to the true oracle $\mathcal{O}$. If, on the other hand, $\hat{Q}(A) \cap Q(B) \subseteq Q(E)$ then there is an oracle $\tilde{\mathcal{O}}$ that is consistent with the sampled view of A and the view of the real B. That is, there is an execution of the protocol with an oracle $\tilde{\mathcal{O}}$ that yields the observed transcript T, a view for B identical to the view of the real B, and a view for A identical to the view generated by E in the current iteration. Perfect

completeness implies that the key k computed by A in this case must match the (actual) key computed by B.

Since B makes at most q_B queries, it follows that there are at most q_B iterations in which E adds an incorrect key to K, and so at least q_B+1 iterations in which E adds the correct key to K. Since E outputs the key that occurs most often, E always outputs the correct key.

4.2 Breaking Key Agreement Relative to $\mathcal{O}, \mathcal{WI}$

Here we prove the main result of this section:

Theorem 4.2. *There is no augmented fully black-box construction of key agreement with perfect completeness from one-way functions.*

The overall structure of the attack is the same as in the previous section, but there are some key differences. Our attack again proceeds by having E repeatedly find a view of A consistent with a transcript T and the oracle queries $Q(E)$ that E has made thus far. Let $Q(A)$ and $Q(B)$ denote the queries of A and B, respectively, in the actual execution of the protocol, and let $\hat{Q}(A)$ denote the queries of A in the view found by E in some iteration. In the previous section we argued that as long as $\hat{Q}(A) \cap Q(B) \subseteq Q(E)$, the key found by E in the given iteration matches the key computed by the real B. This was because, under that condition, there must exist an oracle $\tilde{\mathcal{O}}$ that is consistent with an execution of the protocol in which party A makes queries $\hat{Q}(A)$, party B makes queries $Q(B)$, and the resulting transcript is T. Here, however, that need not be the case. For example, consider a real execution of the protocol in which B makes a query $\mathcal{V}(x,\pi)$ that returns 1, yet B does not make any corresponding query $\mathcal{P}(x,w,\star)=\pi$ with $(x,w) \in R_L$. If E samples a view of A in which $x \notin L$, then there are no oracles $\tilde{O}, \widetilde{\mathcal{WI}}$ consistent with the sampled view of A and the real view of B, but neither does E necessarily learn any new queries in $Q(B)$.

We deal with the above by modifying the attack and changing the proof. First, we modify the attack by having E sample *extended* views of A, which include a view of A along with additional oracle queries used for "book-keeping". Second, rather than showing that, in every iteration, E either learns the correct key or a query in $Q(B)$, we show that, in every iteration, E either learns the correct key or a query in $Q(AB) \stackrel{\text{def}}{=} Q(A) \cup Q(B)$.

An additional subtlety arises due to the possibility that Spoof_i occurs (cf. Lemma 2.4) for some i. In our attack we handle this by guaranteeing that $\mathsf{Spoof} = \cup_i \mathsf{Spoof}_i$ occurs with sufficiently small probability, and showing that the attack is successful whenever Spoof does not occur. (Our proof can be significantly simplified if we make the assumption that $A(1^n)$ and $B(1^n)$ only query their oracles on inputs of length n, however we wish to avoid this assumption.)

Preliminaries: We view $Q(A), Q(B)$, and $Q(E)$ interchangeably as sets of queries and sets of query/answer pairs. We write, e.g., $[\mathcal{P}(x,w,r)=\pi] \in Q(A)$

to denote that A made the query $\mathcal{P}(x, w, r)$ and received the answer π. As usual, we let L denote the set of satisfiable circuits with $\mathcal{O}$-gates.

We assume any key-agreement construction (A, B) has the following normal form: Before a party queries $\mathcal{P}(x, w, r)$, that party also asks all $\mathcal{O}$-queries necessary to check whether $(x, w) \in R_L$; after receiving the result $\pi = \mathcal{P}(x, w, r)$, that party also asks $\mathcal{V}(x, \pi)$. Any key-agreement protocol can be modified to satisfy this condition with only a polynomial blow-up in the number of queries. We let $q = q(n) \geq n$ denote a polynomial upper bound on the combined running time of A and B (and so in particular a bound on the number of queries they make).

Without loss of generality, assume that for any (circuit) $x \in \{0,1\}^n$ and $w \in \{0,1\}^n$, computation of x on input w queries $\mathcal{O}$ at most n times, each time on input of length at most n.

Extended views of A: In our attack, E will repeatedly sample *extended* views of A which include A's view along with some additional oracle queries/answers. We denote an extended view by $(r_A, \mathcal{O}', \mathcal{WI}')$, where r_A are the random coins of A and $\mathcal{O}', \mathcal{WI}'$ are a set of query/answer pairs that includes all those made by A (using coins r_A and the given transcript). E samples only *consistent* extended views, which we define now.

Definition 4.3. *Let $Q = \left(\mathcal{O}', \mathcal{WI}' = (\mathcal{P}', \mathcal{V}')\right)$ be a set of queries/answers. We say it is* consistent *if*

1. *For every query $[\mathcal{P}'(x, w, r) = \pi] \in \mathcal{WI}'$, oracle $\mathcal{O}'$ contains queries/answers sufficient to determine whether $(x, w) \in R_L$. Moreover, if $(x, w) \in R_L$ then $[\mathcal{V}'(x, \pi) = 1] \in \mathcal{WI}'$, while if $(x, w) \notin R_L$ then $[\mathcal{V}'(x, \pi) = 0] \in \mathcal{WI}'$.*
2. *For every query $[\mathcal{V}'(x, \pi) = 1] \in \mathcal{WI}'$, there exist w, r such that $\mathcal{O}'$ contains queries/answers for which $(x, w) \in R_L$ and $[\mathcal{P}'(x, w, r) = \pi] \in \mathcal{WI}'$.*

Let T be a transcript of an execution between $A(1^n)$ and $B(1^n)$, and let $Q(E)$ be a set of queries/answers. We say the extended view $(r_A, \mathcal{O}', \mathcal{WI}')$ is consistent with T and $Q(E)$ *if $\mathcal{O}', \mathcal{WI}'$ is consistent, and also:*

1. *Every query in $Q(E)$ is in $\mathcal{O}', \mathcal{WI}'$, and is answered the same way.*
2. *$A^{\mathcal{O}', \mathcal{WI}'}(1^n; r_A)$, when fed with incoming messages as in T, would generate outgoing messages consistent with T.*

The attack. Let $t = 4 \log q$. First, E queries $\mathcal{O}(x)$ for all x with $|x| \leq t$; queries $\mathcal{P}(x, w, r)$ for all x, w, r with $|x| = |w| = |r| \leq t$; and queries $V(x, \pi)$ for all x, π with $|x| = |\pi|/7 \leq t$. Denote these queries/answers by $Q^*(E)$. The rest of the attack is similar to that of the previous section. E, given a transcript T of an execution of (A, B), initializes $Q(E) = Q^*(E)$ and $K = \emptyset$, and then runs $2q + 1$ iterations of the following:

- *Simulation phase:* E finds an extended view $(r_A, \mathcal{O}', \mathcal{WI}')$ consistent with T and $Q(E)$, with $\mathcal{O}', \mathcal{WI}'$ of size at most $|Q(E)| + q$. (If no such extended view exists, E aborts.) Let k be the key computed by A in this view. E adds k to K.

– *Update phase:* E makes all queries in $(\mathcal{O}' \cup \mathcal{WI}') \setminus Q(E)$ to the true oracles $\mathcal{O}, \mathcal{WI}$. For any queries $[\mathcal{P}'(x,w,r) = \pi]$ just made, E also makes any $\mathcal{O}$ queries needed to determine whether $(x,w) \in R_L$, as well as the query $\mathcal{V}(x,\pi)$. All the resulting query/answer pairs are added to $Q(E)$.

Following the above, E has a multiset K of $2q+1$ possible keys. E outputs the majority value in K.

Analysis. In pre-processing, E makes polynomially many queries. In each iteration of the attack, E makes at most $q + q(q+1) \leq 3q^2$ queries: there are at most q queries in $(\mathcal{O}' \cup \mathcal{WI}') \setminus Q(E)$, and for each such query of the form $[\mathcal{P}'(x,w,r) = \pi]$ we have $|x| \leq q$ and so at most q queries are needed to check whether $(x,w) \in R_L$. Thus, E makes at most $7q^3$ queries after the pre-processing.

For any i, define Spoof_i (cf. Lemma 2.4) to be the event that there is a query $[\mathcal{V}_i(x,\pi) = 1] \in Q(A) \cup Q(B)$, yet there is no query

$$[\mathcal{P}_i(x,w,\star) = \pi] \in Q(A) \cup Q(B) \cup Q^*(E)$$

with $(x,w) \in R_L$. Let $\mathsf{Spoof} = \bigvee_i \mathsf{Spoof}_i$. We claim that Spoof occurs with probability at most $1/4$. Indeed, by construction Spoof_i cannot occur for $i \leq t$, and (by Lemma 2.4 and a union bound) $\Pr[\bigvee_{i>t} \mathsf{Spoof}_i] \leq 1/8$.

Define Spoof' to be the event that, at some point during the attack, E queries $\mathcal{V}(x,\pi) = 1$ to the real oracle, but there was no previous query $[\mathcal{P}_i(x,w,\star) = \pi]$ made by A, B, or E with $(x,w) \in R_L$. By construction, this can only possibly occur if $|x| > 4 \log q$. Since E makes at most $7q^3$ queries after the pre-processing stage, however, Spoof' occurs with probability at most $1/8$.

In the rest of the analysis, we show that as long as neither Spoof nor Spoof' occur, E outputs the key computed by A and B. This suffices, since then E finds the shared key with probability at least $3/4$ overall. As in the previous section, then, the following lemma will prove Theorem 4.2:

Lemma 4.4. *Let k denote the actual key computed by A and B in an execution of the protocol, and assume neither Spoof nor Spoof' occur. Then E does not abort, and in each iteration of the attack either E adds k to K, or E adds to $Q(E)$ one of the queries made by A or B in the real execution.*

Proof. Let $Q(AB) \stackrel{\text{def}}{=} Q(A) \cup Q(B)$ denote the queries/answers made/received by A or B in the real execution. We first show that E never aborts. Say $Q(E)$ is consistent at the beginning of some iteration; this is true by construction in the first iteration. Since Spoof did not occur, a consistent, extended view is given by letting $(\mathcal{O}', \mathcal{WI}') = Q(E) \cup Q(AB)$, which is of size at most $|Q(E)| + q$. Moreover, regardless of what consistent, extended view is actually sampled by E, the new set $Q(E)$ defined at the end of the iteration is consistent unless Spoof' occurs.

We now prove the rest of the lemma. Let $(r_A, \mathcal{O}', \mathcal{WI}')$ be the consistent, extended view chosen by E in some iteration. We define three events, and show:

– If one of the events occurs, then, in the update phase of that iteration, E adds to $Q(E)$ some query in $Q(AB)$.

– If none of the events occur then there are oracles $\tilde{\mathcal{O}}, \widetilde{\mathcal{WI}}$ that match (i.e., are not inconsistent with) the extended view of A and the real view of B. (Thus, by perfect completeness, E adds the correct key to K in that iteration.)

Before defining the events, we introduce some terminology. Given some set of queries Q, we say Q *fixes* $x \in L$ if either (1) there exists a w and $\mathcal{O}$-queries in Q such that $(x, w) \in R_L$, or (2) there is a query $[\mathcal{V}(x, \star) = 1] \in Q$. We say Q *fixes* $x \notin L$ if for all w there are $\mathcal{O}$-queries in Q such that, regardless of how any of the $\mathcal{O}$-queries not in Q are answered, it holds that $(x, w) \notin R_L$. We define Q *fixes* $(x, w) \in R_L$ and Q *fixes* $(x, w) \notin R_L$ in the obvious way.

We now define the events of interest:

E_1**:** $\mathcal{O}', \mathcal{WI}'$ disagrees with $Q(AB)$ on the answer to some $\mathcal{O}$-, $\mathcal{P}$-, or $\mathcal{V}$-query.

E_2**:** There exists an x such that $Q(AB)$ fixes $x \in L$ but $\mathcal{O}', \mathcal{WI}'$ fixes $x \notin L$, or vice versa.

E_3**:** A $\mathcal{V}'$-query returning 0 in $\mathcal{WI}'$ is "inconsistent" with the $\mathcal{O}, \mathcal{P}$ queries in $Q(AB)$, or vice versa. Formally, one of the following occurs:

- There is a query $[\mathcal{V}'(x, \pi) = 0] \in \mathcal{WI}'$, but $[\mathcal{P}(x, w, \star) = \pi] \in Q(AB)$ and $Q(AB)$ fixes $(x, w) \in R_L$.
- There is a query $[\mathcal{P}'(x, w, \star) = \pi] \in \mathcal{WI}'$ and $\mathcal{O}'$ fixes $(x, w) \in R_L$, but $[\mathcal{V}(x, \pi) = 0] \in Q(AB)$.

Claim 4.5. *If any of E_1, E_2, or E_3 occur in the simulation phase of some iteration, then E learns a new query in $Q(AB)$ in the update phase of that iteration.*

Proof. If E_1 occurs, the claim is immediate. ($Q(E)$ contains the answers of the true oracles, and so can never disagree with $Q(AB)$. So any disagreement between $\mathcal{O}', \mathcal{WI}'$ and $Q(AB)$ must be due to some query in $\mathcal{O}', \mathcal{WI}'$ outside of $Q(E)$.) If E_2 occurs there are several sub-cases to consider:

1. Say $Q(AB)$ fixes $x \in L$, but $\mathcal{O}', \mathcal{WI}'$ fixes $x \notin L$. The second event implies that for all w oracle $\mathcal{O}'$ fixes $(x, w) \notin R_L$. There are two ways the first event can occur:
   - There exists a w such that $Q(AB)$ fixes $(x, w) \in R_L$. In this case there must be an $\mathcal{O}$-query in $Q(AB)$ that is answered inconsistently with some query in $\mathcal{O}'$, and event E_1 has occurred.
   - There is a query $[\mathcal{V}(x, \pi) = 1] \in Q(AB)$ (for some π). Since Spoof has not occurred, this means that for some w, r there is a query $[\mathcal{P}(x, w, r) = \pi]$ in $Q(AB)$ or $Q^*(E)$. Say $[\mathcal{P}(x, w, r) = \pi] \in Q(AB)$. Then by our normal-form assumption, $Q(AB)$ fixes $(x, w) \in R_L$; this, in turn, implies an $\mathcal{O}$-query in $Q(AB)$ inconsistent with $\mathcal{O}'$ (which, recall, fixed $x \notin L$), and so E_1 has occurred.
     On the other hand, say $[\mathcal{P}(x, w, r) = \pi] \in Q^*(E)$. Then, by definition of $Q^*(E)$, the query $[\mathcal{V}(x, \pi) = 1]$ is also in $Q^*(E)$, and $Q^*(E)$ fixes $(x, w) \in R_L$. But since any queries in $\mathcal{O}'$ must agree with the corresponding $\mathcal{O}$-queries in $Q^*(E)$, this cannot happen.

2. Say $\mathcal{O}', \mathcal{WI}'$ fixes $x \in L$, but $Q(AB)$ fixes $x \notin L$. The second event implies that for all w we have that $Q(AB)$ fixes $(x, w) \notin R_L$. There are two ways the first event can occur:
   - There exists a w for which $\mathcal{O}'$ fixes $(x, w) \in R_L$. In this case there is an $\mathcal{O}$-query in $Q(AB)$ that is answered inconsistently with some query in $\mathcal{O}'$, and event E_1 has occurred.
   - There is a query $[\mathcal{V}(x, \pi) = 1] \in \mathcal{WI}'$ for some π. By definition of consistency, there exists w such that $\mathcal{O}'$ fixes $(x, w) \in R_L$. Then there must be an $\mathcal{O}$-query in $Q(AB)$ that is answered inconsistently with $\mathcal{O}'$, and so E_1 has occurred.

Finally, we turn to E_3. Here there are two sub-cases:

1. Say $[\mathcal{V}'(x, \pi) = 0] \in \mathcal{WI}'$, but $[\mathcal{P}(x, w, \star) = \pi] \in Q(AB)$ and furthermore $Q(AB)$ fixes $(x, w) \in R_L$. Because of our normal-form assumption, $[\mathcal{V}(x, \pi) = 1] \in Q(AB)$. Thus there is a $\mathcal{V}$-query in $Q(AB)$ that is answered inconsistently with $\mathcal{WI}'$ and so E_1 has occurred.
2. Say $[\mathcal{P}'(x, w, \star) = \pi] \in \mathcal{WI}'$ and $\mathcal{O}'$ fixes $(x, w) \in R_L$, but we have $[\mathcal{V}(x, \pi) = 0] \in Q(AB)$. By definition of consistency, $[\mathcal{V}(x, \pi) = 1] \in \mathcal{WI}'$. Thus there is a $\mathcal{V}$-query in $Q(AB)$ that is answered inconsistently with $\mathcal{WI}'$, and so E_1 has occurred.

This concludes the proof of Claim 4.5.

To complete the proof of the lemma, we show that if none of E_1, E_2, or E_3 occur, there exist oracles $\tilde{\mathcal{O}}, \widetilde{\mathcal{WI}}$ (in the support of the distribution from Section 2.1) that match (i.e., do not disagree with) $\mathcal{O}', \mathcal{WI}'$, and $Q(AB)$. This means there is an execution of the protocol with oracles $\tilde{\mathcal{O}}, \widetilde{\mathcal{WI}}$ that yields a view for B identical to the view of the real B, and a view for A identical to the view of A in the extended view sampled by E. Perfect completeness implies that the key k computed by A in that case must match the (actual) key computed by B, as we needed to show.

We construct $\widetilde{\mathcal{O}}, \widetilde{\mathcal{WI}}$ as follows. First, answer all queries in $\mathcal{O}', \mathcal{WI}'$, and $Q(AB)$ as answered by those oracles; if E_1 does not occur, this is well-defined as there is no conflict. Answer all other queries in $\tilde{\mathcal{O}}$ arbitrarily. Note that if $\mathcal{O}', \mathcal{WI}', Q(AB)$ fixes $x \in L$ then so does $\tilde{\mathcal{O}}$, and similarly if $\mathcal{O}', \mathcal{WI}', Q(AB)$ fixes $x \notin L$. Note also that with $\tilde{\mathcal{O}}$ fixed, so are $\tilde{L}$ and $\tilde{R}_L$.

For $\tilde{\mathcal{P}}$, proceed as follows. Recall that all $\tilde{\mathcal{P}}_i$ queries for $i \leq t = 4 \log q$ were made by E during pre-processing and so are already fixed. Any other unassigned query $\tilde{\mathcal{P}}(x, w, r)$ with $|x| > t$ is defined as follows:

- If $(x, w) \notin \tilde{R}_L$, the query is answered arbitrarily.
- If $(x, w) \in \tilde{R}_L$, let $\pi^* \in \{0,1\}^{7|x|}$ be such that $\mathcal{V}(x, \pi^*)$ is not in $\mathcal{WI}'$ or $Q(AB)$. (There must exist such a π^*, by the bound on the number of queries in these sets.) Set $\tilde{\mathcal{P}}(x, w, r) = \pi^*$.

With the $\tilde{\mathcal{O}}$ and $\tilde{\mathcal{P}}$ queries fixed, oracle $\tilde{\mathcal{V}}$ is set as in Section 2.1.

We show that $\widetilde{\mathcal{O}}, \widetilde{\mathcal{WI}}$ match (i.e., do not disagree with) $\mathcal{O}', \mathcal{WI}'$, and $Q(AB)$. By construction, the only possible conflict can be between $\tilde{\mathcal{V}}$ and some $\mathcal{V}$-query in $\mathcal{WI}'$ or $Q(AB)$. No such conflict is possible:

1. Say $[\mathcal{V}(x,\pi)=1] \in \mathcal{WI}'$ for some x, π. Then by definition of consistency, there exist w, r such that $\mathcal{O}'$ fixes $(x,w) \in R_L$, and $[\mathcal{P}(x,w,r)=\pi] \in \mathcal{WI}'$. But then $(x,w) \in \tilde{R}_L$ and $\tilde{\mathcal{P}}(x,w,r)=\pi$, and so $\tilde{\mathcal{V}}(x,\pi)=1$.
2. Say $[\mathcal{V}(x,\pi)=1] \in Q(AB)$ for some x, π. Since Spoof does not occur, there exist w, r such that $\mathcal{O}' \cup Q(AB)$ fixes $(x,w) \in R_L$, and $[\mathcal{P}(x,w,r)=\pi] \in \mathcal{WI}' \cup Q(AB)$. But then $(x,w) \in \tilde{R}_L$ and $\tilde{\mathcal{P}}(x,w,r)=\pi$, and so $\tilde{\mathcal{V}}(x,\pi)=1$.
3. Say $[\mathcal{V}(x,\pi)=0] \in \mathcal{WI}' \cup Q(AB)$ for some x, π. If $x \notin \tilde{L}$ then $\tilde{\mathcal{V}}(x,\pi)=0$ also. If $x \in \tilde{L}$, there is an inconsistency only if there is some w with $\tilde{\mathcal{P}}(x,w,\star)=\pi$ and $(x,w) \in \tilde{R}_L$. Note $\tilde{\mathcal{P}}(x,w,\star)=\pi$ can only occur if $[\mathcal{P}(x,w,\star)=\pi] \in \mathcal{WI}' \cup Q(AB)$, but in that case (since $[\mathcal{V}(x,\pi)=0] \in \mathcal{WI}' \cup Q(AB)$ and E_3 did not occur) either $\mathcal{O}'$ or $Q(AB)$ fix $(x,w) \notin R_L$, and hence $(x,w) \notin \tilde{R}_L$ either.

This completes the proof of Lemma 4.4.

5 Impossibility for Statistically-Hiding Commitments

We show that the impossibility results of Haitner et al. [12, 13] for statistically-hiding commitment schemes can be strengthened to hold even within our new framework. (Our results here do not require perfect completeness.)

Theorem 5.1. *Any augmented fully black-box construction of a statistically-hiding bit-commitment scheme from trapdoor permutations over $\{0,1\}^n$ has an $\Omega(n/\log n)$-round commit stage.*

Theorem 5.2. *Any augmented fully black-box construction of a statistically-hiding bit-commitment scheme from trapdoor permutations over $\{0,1\}^n$ requires the sender to communicate $\Omega(n)$ bits to the receiver during the commit stage.*

Note that in the above theorems we consider constructions which invoke only trapdoor permutations over n bits, where n is the security parameter. In fact, when considering constructions which may invoke the trapdoor permutations over smaller domains, better upper bounds are known. In particular, it is possible to apply the scheme of Naor et al. [17] using a one-way permutation over n^ϵ bits, which results in a statistically-hiding commitment scheme with an $O(n^\epsilon)$-round commit phase. As already discussed by Haitner et al. [12] this issue is not unique to our setting, but arises in essentially any study of the *efficiency* of cryptographic reductions. The common approach for addressing this issue is by restricting the class of constructions (as in the statements of our theorems); we refer the reader to [12] for a less restrictive approach.

Due to space limitations the proofs of Theorems 5.1 and 5.2 are provided in the full version of this work, and here we only give a high-level overview. We

consider a set of oracles $\Gamma = (\mathcal{O}, \mathcal{P}, \mathcal{V}, \mathsf{Sam})$, and prove that the following hold with high probability[1]:

1. $\mathcal{O}$ is a collection of trapdoor permutations relative to Γ.
2. $(\mathcal{P}, \mathcal{V})$ is a WI proof system for $\mathcal{NP}^{\mathcal{O}}$ relative to Γ.
3. Any statistically-hiding bit-commitment scheme in which the sender and receiver have oracle access to $(\mathcal{O}, \mathcal{P}, \mathcal{V})$ can be broken using Γ. The efficiency and success probability of the attack depend on the round complexity or communication complexity of the commitment scheme.

This suffices because any (augmented) fully black-box construction is also *relativizing* [19].

The oracle Sam is the interactive collision-finding oracle of Haitner et al. [12]. In its most basic and non-interactive form, this oracle takes as input a circuit C, and outputs a random pair of inputs (w, w') such that $C(w) = C(w')$. Relative to such an oracle there are no collision-resistant hash functions [21] or 2-move statistically-hiding commitment schemes [8]. Moreover [21], one-way functions exist relative to Sam. This oracle was generalized by Haitner et al. to an interactive setting: Sam takes as input a circuit C and a "target" value z, and outputs a random input w such that $C(w) = z$. Haitner et al. had to force various restrictions on Sam such that one-way functions continue to exist, yet Sam remains sufficiently powerful to break the binding of (interactive) statistically-hiding commitment schemes.

In our setting, where we also consider a WI oracle $(\mathcal{P}, \mathcal{V})$, the argument that Sam can be used to break statistically-hiding commitment schemes is essentially identical to the corresponding argument of Haitner et al. [12, 13]. The challenging part (in which our proof differs from that of Haitner et al.), lies in showing that one-way functions (and, more generally, that trapdoor permutations) still exist relative to Sam, and that $(\mathcal{P}, \mathcal{V})$ is still witness indistinguishable.

The proofs of these properties are more subtle than the corresponding proofs in Section 2. In that section we relied on the fact that any efficient algorithm can issue only a polynomial number of queries to $\mathcal{O}$ and $\mathcal{P}$. Here, however, when considering also the oracle Sam, this is no longer true: although an efficient algorithm with access to Γ may issue only a polynomial number of *direct queries* to $\mathcal{O}$, $\mathcal{P}$, and Sam, the oracles $\mathcal{O}$ and $\mathcal{P}$ may actually be *indirectly queried* an exponential number of times by Sam, and the previous arguments no longer hold.

To circumvent this and several other similar difficulties, we extend the proof of Haitner et al. [12] that manages to distinguish between the amount of "useful information" that is obtained by direct and indirect queries, and uses information-theoretic compression arguments that are oblivious to the (possibly exponential) number of indirect queries. The main difficulty in our setting, when compared to that of [12], is that we need to deal also with the oracles $\mathcal{P}$ and $\mathcal{V}$. Note that $\mathcal{P}$ is simply a random function (and thus can be treated as in [12]), but $\mathcal{V}$ has structure. Technically, proving that $\mathcal{O}$ is one-way is very similar to the

[1] We prove our statements with respect to a distribution over oracles. As in Lemma 2.6, we can also reverse the order of quantifiers and fix a specific oracle.

corresponding proof in [12], since when compressing the description of $\mathcal{O}$ we are granted unbounded access to $\mathcal{P}$ and $\mathcal{V}$, and this enables us to perfectly simulate their behavior. The main difference is in proving that $(\mathcal{P}, \mathcal{V})$ is a WI proof system, and due to the structure of $\mathcal{V}$ this requires us to refine and adjust the compression arguments from [12] for arguing that $\mathcal{V}$ does not reveal too much "useful information" on $\mathcal{P}$, and can be simulated while compressing $\mathcal{P}$.

Finally, we note that although we prove our impossibility results for non-interactive *witness-indistinguishable* proof systems, our results immediately extend to non-interactive *zero-knowledge* proof systems (the main difference is in allowing the sender and receiver access to a common reference string). This follows from the fact that our impossibility results hold even for commitment schemes in which the hiding property is assumed to hold only with respect to the honest receiver (exactly as in [12, 13]). Therefore, in such a case the receiver can choose a common random string that transforms a witness-indistinguishable proof system into a zero-knowledge proof system as in Section 2.2. Specifically, showing that $\mathcal{O}$ is one-way relative to Γ implies the existence of a pseudorandom generator, and therefore the transformation in Section 2.2 can be carried out by having the receiver sample a uniform random string and send it to the sender in the first round.

Acknowledgments

We thank the anonymous referees for their extensive comments, and Dominique Unruh for a thorough proofreading of our results and several useful discussions.

References

1. B. Applebaum, Y. Ishai, and E. Kushilevitz. Computationally private randomizing polynomials and their applications. *Computational Complexity*, 15(2):115–162, 2006.
2. B. Barak and M. Mahmoody-Ghidary. Merkle puzzles are optimal: An $O(n^2)$-query attack on key exchange from a random oracle. In *Advances in Cryptology — Crypto 2009*, volume 5677 of *LNCS*, pages 374–390. Springer, 2009.
3. D. Beaver. Correlated pseudorandomness and the complexity of private computations. In *28th Annual ACM Symposium on Theory of Computing (STOC)*, pages 479–488. ACM Press, 1996.
4. M. Bellare and S. Goldwasser. New paradigms for digital signatures and message authentication based on non-interative zero knowledge proofs. In *Advances in Cryptology — Crypto '89*, volume 435 of *LNCS*, pages 194–211. Springer, 1990.
5. D. Dolev, C. Dwork, and M. Naor. Nonmalleable cryptography. *SIAM Journal on Computing*, 30(2):391–437, 2000.
6. U. Feige, A. Fiat, and A. Shamir. Zero-knowledge proofs of identity. *Journal of Cryptology*, 1(2):77–94, 1988.
7. U. Feige, D. Lapidot, and A. Shamir. Multiple non-interactive zero knowledge proofs under general assumptions. *SIAM Journal on Computing*, 29(1):1–28, 1999.

8. M. Fischlin. On the impossibility of constructing non-interactive statistically-secret protocols from any trapdoor one-way function. In *Cryptographers' Track — RSA 2002*, volume 2271 of *LNCS*, pages 79–95. Springer, 2002.
9. M. Fischlin. Round-optimal composable blind signatures in the common reference string model. In *Advances in Cryptology — Crypto 2006*, volume 4117 of *LNCS*, pages 60–77. Springer, 2006.
10. R. Gennaro, Y. Gertner, J. Katz, and L. Trevisan. Bounds on the efficiency of generic cryptographic constructions. *SIAM Journal on Computing*, 35(1):217–246, 2005.
11. Y. Gertner, T. Malkin, and S. Myers. Towards a separation of semantic and CCA security for public key encryption. In *4th Theory of Cryptography Conference — TCC 2007*, volume 4392 of *LNCS*, pages 434–455. Springer, 2007.
12. I. Haitner, J. J. Hoch, O. Reingold, and G. Segev. Finding collisions in interactive protocols — a tight lower bound on the round complexity of statistically-hiding commitments. In *48th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 669–679. IEEE, 2007.
13. I. Haitner, J. J. Hoch, and G. Segev. A linear lower bound on the communication complexity of single-server private information retrieval. In *5th Theory of Cryptography Conference — TCC 2008*, volume 4948 of *LNCS*, pages 445–464. Springer, 2008.
14. 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.
15. R. Impagliazzo and S. Rudich. Limits on the provable consequences of one-way permutations. In *21st Annual ACM Symposium on Theory of Computing (STOC)*, pages 44–61. ACM Press, 1989.
16. Y. Lindell. A simpler construction of CCA2-secure public-key encryption under general assumptions. In *Advances in Cryptology — Eurocrypt 2003*, volume 2656 of *LNCS*, pages 241–254. Springer, 2003.
17. M. Naor, R. Ostrovsky, R. Venkatesan, and M. Yung. Perfect zero-knowledge arguments for NP using any one-way permutation. *Journal of Cryptology*, 11(2):87–108, 1998.
18. M. Naor and M. Yung. Public-key cryptosystems provably secure against chosen ciphertext attacks. In *22nd Annual ACM Symposium on Theory of Computing (STOC)*, pages 427–437. ACM Press, 1990.
19. O. Reingold, L. Trevisan, and S. P. Vadhan. Notions of reducibility between cryptographic primitives. In *1st Theory of Cryptography Conference — TCC 2004*, volume 2951 of *LNCS*, pages 1–20. Springer, 2004.
20. A. Sahai. Non-malleable non-interactive zero knowledge and adaptive chosen-ciphertext security. In *40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 543–553. IEEE, 1999.
21. D. R. Simon. Finding collisions on a one-way street: Can secure hash functions be based on general assumptions? In *Advances in Cryptology — Eurocrypt '98*, volume 1403 of *LNCS*, pages 334–345. Springer, 1998.