On the (Im)Possibility of Arthur-Merlin Witness Hiding Protocols

Iftach Haitner[1⋆], Alon Rosen[2⋆⋆], and Ronen Shaltiel[3⋆⋆⋆]

[1] Microsoft Research, New England Campus.
iftach@microsoft.com
[2] Herzliya Interdisciplinary Center, Herzliya, Israel.
alon.rosen@idc.ac.il
[3] University of Haifa.
ronen@cs.haifa.ac.il

Abstract. The concept of *witness-hiding* suggested by Feige and Shamir is a natural relaxation of zero-knowledge. In this paper we identify languages and distributions for which many known constant-round public-coin protocols with negligible soundness cannot be shown to be witness-hiding using black-box techniques. One particular consequence of our results is that parallel repetition of either 3-Colorability or Hamiltonicity cannot be shown to be witness hiding with respect to some probability distribution over the inputs assuming that:

1. the distribution assigns positive probability only to instances with *exactly one* witness.
2. Polynomial size circuits cannot find a witness with noticeable probability on a random input chosen according to the distribution.
3. The proof of security relies on a black-box reduction that is *independent* of the choice of the commitment scheme used in the protocol.

These impossibility results conceptually match results of Feige and Shamir that use such black-box reductions to show that parallel repetition of 3-Colorability or Hamiltonicity *is* witness-hiding for distributions with "two independent witnesses".

We also consider black-box reductions for parallel repetition of 3-Colorability or Hamiltonicity that *depend* on a specific implementation of the commitment scheme. While we cannot rule out such reductions completely, we show that "natural reductions" cannot bypass the limitations above.

Our proofs use techniques developed by Goldreich and Krawczyk for the case of zero knowledge. The setup of witness-hiding, however, presents new technical and conceptual difficulties that do not arise in the zero-knowledge setting. The high level idea is that if a black-box reduction establishes the witness-hiding property for a protocol, and the protocol also happens to be a proof of knowledge, then this latter property can be actually used "against the reduction" to find witnesses unconditionally.

Keywords: Zero-Knowledge, Witness-Hiding, Arthur Merlin protocols, Black-box reductions

⋆ Part of this work performed while at the University of Haifa.
⋆⋆ Research supported by BSF grant 2006317.
⋆⋆⋆ Research supported by BSF grant 2004329 and ISF grant 686/07.

1 Introduction

In a proof the *prover* tries to convince the *verifier* that a certain statement is true. The basic requirements are *completeness* and *soundness*. The former means that the prover is always able to convince the verifier in the validity of a true statement, while the latter means that the prover is not able to convince the verifier in the validity of a false statement.

In cryptography, the statement typically belongs to NP, and the proof is required to maintain the prover's "privacy". As a consequence, the proof is interactive and randomized, and the verifier only gets statistical confidence in the validity of the statement.

The privacy requirement usually refers to some information about the NP-witness used by the prover. Given the difficulty in capturing what exactly is the information that needs to be hidden, the tendency is to be conservative. This gives rise to the notion of *zero-knowledge* proofs: protocols that do not reveal *anything* beyond the validity of the statement being proved [15].

1.1 Zero Knowledge

Zero-knowledge (ZK) proofs are usually constructed using smaller "atomic" ZK protocols as a building block. The typical atomic protocol is "public coin", requires 3 rounds of interaction, and may convince the verifier in the validity of a false statement with constant probability.[4] Well known examples are the protocols for Quadratic Residuosity [15], 3-Colorability [13], and Hamiltonicity [5].

In order to gain higher statistical confidence in the validity of the statement, the verifier requests to repeat the execution of the atomic ZK protocol multiple times independently. To retain the zero-knowledge property of the atomic protocol, the verifier requests that the repetitions be conducted *sequentially* [14]. This results in high round complexity, and is highly undesirable.

An alternative way to increase the verifier's confidence (preferable in terms of round complexity) is to repeat the sub-protocol in *parallel*. Demonstrating that parallel repetition of the atomic protocols is zero-knowledge, however, appears to be a challenging task. Indeed, as shown by Goldreich and Krawczyk only trivial languages have a black-box zero-knowledge constant-round public-coin proof with negligible soundness error [12].[5]

The Goldreich-Krawczyk impossibility result can be bypassed by considering private-coin protocols [12], or by employing non black-box simulation techniques [3]. Nevertheless, it is still interesting to ask whether black-box techniques can be used to establish security of public-coin protocols. First of all, public-coin

[4] A protocol is public-coin if the verifier's messages consist of his own random coins. Constant-round, public-coin proofs are sometimes referred to as *Arthur-Merlin proofs* (AM) in the literature (cf., [2])

[5] Black-box zero-knowledge essentially means that when establishing the zero-knowledge property of the protocol, the protocol designer is restricted to only observing the "input-output" behavior of a given malicious verifier.

protocols tend to be simpler than private-coin ones, and are easier to work with when used as sub-protocols. Secondly, current non black-box techniques are only known to achieve soundness against computationally bounded provers, whereas the "atomic" ZK protocols retain their soundness even in face of a computationally unbounded prover. Thirdly, many of the known public-coin protocols require only 3 rounds of communication, whereas the best private-coin and non-black protocols require 5 and 7 rounds respectively. Finally, when available, black-box techniques are preferable over their "non-black-box" counterparts, mainly because they offer a better tradeoff of security vs. efficiency.

1.2 Witness Indistinguishability

Despite the failure in establishing the zero-knowledge property of constant-round public-coin protocols with negligible soundness, and in particular of the parallel repetition of "atomic" ZK, there is still no evidence that these specific protocols are insecure. This suggests an alternative approach: identify a weaker (yet meaningful) security property and prove that it is satisfied by the protocols.

Feige and Shamir define a protocol to be *witness-indistinguishable* (WI) if the verifier cannot identify the witness that was actually used by the prover in the interaction [9]. Witness-indistinguishability is implied by zero knowledge. Unlike zero-knowledge, however, witness indistinguishability is preserved under parallel repetition. As a consequence, repeating atomic ZK protocols in parallel results in a 3-round WI public-coin protocol with negligible soundness error.

Witness-indistinguishability has turned out to be highly applicable as a building block for higher level protocols. It is sometimes unclear, however, what exactly is hidden by such protocols. On the one hand, WI gives no security guarantee in case that the statement has only one witness associated with it. On the other hand, if a statement has at least two independent witnesses, then any WI protocol for that statement does not reveal the witness being used by the prover in the interaction [9].[6]

[6] The precise formulation of this implication is somewhat technical to state. See [11], Sec 4.6.3.2 for more details. Let us briefly review the argument in a special case. Let f be a one way function and consider the distribution X defined by $x = f(s)$ for a uniformly chosen s. Now consider the distribution X' that consists of two independent copies of X and let $L' = \{x_1, x_2 : \exists i, s_i \text{ s.t. } f(s_i) = x_i\}$. Note that $L' \in NP$ and L' has at least two witnesses on any input in the support of X'. [9] show that a WI proof for L' is hiding witnesses for the distribution X'. Loosely speaking, this follows because if a verifier V^* finds witnesses on X' then V^* can be used to invert f as follows: Given x chosen from X, one can sample another x' from X together with a witness s'. We then set (x_1, x_2) to be a random ordering of (x, x') and prove that the pair $(x_1, x_2) \in L$ to the verifier using the witness s'. We have assumed that the verifier produces a preimage for one of the two x's with noticeable probability. Since the proof is WI the verifier "does not know" which of the two x's was used, V^* can be used to invert the one-way function with noticeable probability.

1.3 Witness Hiding

Motivated by the above observation, Feige and Shamir put forward the notion of *witness-hiding* [9]. Loosely speaking, a protocol is said to be witness-hiding (WH) if at the end of the protocol the verifier cannot compute any new witness that he did not know before the protocol began. This is a natural security requirement, and can replace zero-knowledge in many cryptographic protocols.

Clearly, any ZK protocol is also WH, for any distribution of instances and for any efficiently computable function of the witness. The converse, however, is not true in general.[7] A well known question in this context is whether the parallel repetition of any of the "classical" 3-round protocols hides "interesting" functions of the witness for "interesting" distributions on the instances.

The results of [9] exhibit languages and distributions for which WI protocols are also WH, and thus provide an example where the answer is affirmative. In particular, this is an example where black-box techniques give 3-round public-coin WH protocols whereas, by the results of [12], black-box techniques cannot give such efficient ZK protocols. We remark that WI is not sufficiently strong to *always* imply WH. For example, WI is meaningless for languages that have only one witness, whereas WH is not.

1.4 Our Contributions

The main question we investigate is for what choices of languages and distributions there exist constant-round public-coin WH protocols with negligible soundness. We identify settings in which black-box techniques *cannot* establish the WH hiding property of such protocols. The precise model and results are described below. Before going into these details let us describe some consequences: Suppose that L is a non-trivial language in NP (meaning that $L \notin \text{BPP}$) and suppose that any input $x \in L$ has exactly one witness. Let X be an arbitrary distribution over inputs in L that is hard in the sense that no polynomial size circuit can produce a witness when given x sampled according to X with noticeable probability. We show that:

- It is impossible to show that parallel repetition of 3-colorability or Hamiltonicity is WH for X using certain black-box techniques. This result stands in contrast to the case where L has two or more witnesses, as in this case there exist distributions X for which there are black-box WH proofs (of the type we ruled out above) for the these protocols [9].
- It is impossible to show that generic constructions of Zaps given in [7] are WH for X using black-box techniques.

A consequence of our results is that there exist pairs of indistinguishable distributions for which parallel repetition of 3-colorability (or Hamiltonicity) cannot be shown to have strong witness indistinguishability using certain black-box techniques. The precise details appear in the full version.

[7] E.g., 2-round WI protocols (ZAPs) [7] cannot be ZK, but are WH in some special cases. E.g., the case that the statement has "two independent witnesses".

Another consequence concerns a recent paper by Pass [17] that, following earlier work [10, 6, 1], investigates the possibility of constructing a one way function whose inversion task is at least as hard as deciding some NP complete language (via Turing reductions). Pass shows a relationship between the existence of a Turing reduction as above and the existence of constant-round, negligible soundness public-coin interactive proofs for SAT that hide the bits of the satisfying assignment. The main point in our context is that Pass's relationship requires that the WH property is established using *black-box techniques*. Our results provide limitations on this approach. The precise details appear in the full version.

1.5 The Notion of Black-Box Witness Hiding

We now explain what we mean by "black-box techniques". For a precise description of the model the reader is referred to Sections 2,3. Our definitions of "black-box WH" follow the framework of "black-box ZK" as defined in [12].

Loosely speaking, the definition of black-box WH requires that the WH property of the protocol is established using a reduction R (the reduction can be thought of as the "black-box simulator" in black-box ZK) that satisfies the following property: When R is given oracle access to a cheating verifier V^* that learns the witness following an interaction with the prover P, and an input x sampled from the distribution X, then either R is able to learn the witness with noticeable probability, or R is able to violate the security assumption on which the protocol is based. We consider several flavors of reductions.

Fully vs. weakly black-box reductions We distinguish between two kinds reductions depending on whether the reductions relies on a generic security assumption, such as "there exist one-way functions" or "there exist bit-commitment schemes", or on a specific security assumption, such as "factoring is a one-way function". To illustrate this distinction consider parallel repetition of the classical protocols for 3-Colorability [15] or Hamiltonicity [5]. These protocols require a bit-commitment scheme and can be seen as generic constructions of proof systems that can use *any* bit-commitment scheme. When considering black-box reductions for showing that these protocols are WH, we distinguish between two cases: A ***fully-black-box reduction*** is oblivious to the choice of commitment scheme and should work for any choice of commitment scheme. This is modeled by giving the reduction oracle access to the commitment scheme and requiring that the reduction works for any implementation of commitment schemes. A ***weakly-black-box reduction*** may be tailored for a specific implementation of the commitment scheme that relies on the hardness of a specific function (e.g., we can consider the protocol for 3-Colorability when implemented using Blum's commitment instantiated with Discrete Log). Naturally, it is more difficult to rule out weakly-black-box reductions than fully black box ones.[8]

[8] Since in all the reductions considered in this paper access the cheating verifier (e.g., the adversary) as a "black box", the above definition of fully-black-box reduction coincides with the standard use of this notion (cf., [18]). In our definition of weakly-black-box reduction, however, the reduction treats the adversary as a black-box and

Oblivious versus Tailored reductions. Recall that the witness-hiding property is defined with respect to some distribution X on inputs in L. A reduction R may be ***tailored*** for a specific distribution X. Alternatively, it may be ***oblivious*** and work for any distribution X over inputs in L.

To illustrate this point note that in the case of ZK, a black-box simulator is an oblivious reduction (as the simulator is able to simulate a transcript of $(P, V^*)(x)$ on every input $x \in L$ and thus on every distribution over such inputs). In the case of witness-hiding the reduction of [9] is an example of a black-box reduction that is tailored to the specific distribution. Specifically, this reduction *critically relies* on the ability to query V^* on inputs x' that are different than the input x given to the reduction. It is easy to show that oblivious reductions (and in particular black-box simulators) do not benefit form such behavior, as V^* may be chosen as a function of x and refuse to answer in interactions on inputs $x' \neq x$. In contrast, in the setup of witness-hiding the verifier V^* must agree to participate in the protocol on a noticeable fraction of inputs x in the support of X in order to break the witness-hiding property with noticeable probability.

The fact that tailored reductions can benefit from querying V^* on many different inputs is a new consideration that does not come up in the setting of ZK. The main technical difficulty dealt with in this paper is the development of techniques that handle such reductions.

Embedding reductions. For tailored reductions we say that R is ***non-embedding*** if for every pair of different inputs, if R queries V^* on both inputs then all queries for one input are made before the first query to the other input. A reduction is *embedding* if it does not obey the previous requirement.

Reductions in the literature. The reductions that establish the ZK property of 3-Colorability and Hamiltonicity are fully-black-box oblivious reductions. In fact, as explained earlier all black-box simulators that establish ZK are *oblivious* reductions. There are examples in the literature where the WH property is established using *tailored* reductions (e.g. the reduction of [9] that we sketched earlier that is a tailored and fully-black-box). However, to the best of our knowledge all the reductions in the literature are *non embedding*.

1.6 Statement of Our Results

We now state our results more precisely and explain which kind of black-box reductions we can rule out. Recall that WH is defined with respect to a language $L \in \mathrm{NP}$ and a distribution X over instances in L that is hard in the following sense. No polynomial-size circuit can produce a witness when given x sampled according to X with noticeable probability. We are assuming that $L \notin \mathrm{BPP}$ and that every $x \in L$ has exactly one witness (0therwise WH may follow from WI).[9]

---

treats the "hardness assumption" arbitrarily. This is with contrast to the standard definition of weakly-black-box reduction, where the reduction treats the adversary arbitrarily and treats the hardness assumption as a black box.

[9] We can relax this assumption and consider protocols where the goal is to hide some specific function g of the witness. We refer to g as a "feature" of the witness. We

Our results apply for any constant-round public-coin protocol that has negligible soundness. We now describe our precise results for various kinds of reductions:

Oblivious reductions: We show that oblivious reductions cannot be used to establish black-box WH even if they are weakly-black-box. This is a simple extension of the lower bounds of [12] on black-box ZK. The precise formulation of this result appears in Theorem 3.1.

Tailored non-embedding reductions: We show that tailored reductions that are non-embedding cannot be used to establish WH of protocols that are proofs of knowledge. This result also applies for weakly-black-box reductions. The precise formulation of this result appears in Theorem 3.2. For this result we need to develop new techniques that can handle tailored reductions.

Recall that parallel repetition of 3-Colorability or Hamiltonicity *are* proofs of knowledge and therefore we obtain results on these specific protocols. As we can handle weakly-black-box reductions, these results apply to *any* implementation of these protocols using any choice of commitment schemes.

Embedding fully-black-box reductions: While we do not know how to handle embedding reductions in general, we can handle embedding reductions that are fully-black-box for protocols with an additional property, which we refer to as TKE for "Transcript Knowledge Extractor". Loosely speaking, such protocols have the property that for any prover P^* that convinces V on some input x with probability that is larger than the soundness of the protocol, V can learn a witness for x at the end of the interaction assuming it can break the security assumption on which the protocol is based. We elaborate on this property below (a precise definition appears in Section 3.4).

What we show is that such protocols cannot have a fully-black-box reduction even if the reduction is tailored and embedding. The precise formulation of this result appears in Theorem 3.3. Many generic constructions in the literature of interactive proofs for NP-complete languages have the TKE property. In particular, the protocols for parallel repetition of 3-colorability or Hamiltonicity are fully-black-box (in the sense that they can use any bit-commitment scheme) and have the TKE property (in the sense that a verifier that can break the commitment scheme can learn the witness). It follows that these protocols cannot have fully-black-box reductions even if the reductions are tailored and embedding.

Another interesting case is that of Zaps [7]. These are 2-round WI proofs. Generic constructions of Zaps are not known to be proofs of knowledge (and therefore the lower bound in the previous item does not apply). Nevertheless, we observe that the generic constructions of Zaps in [7] have the TKE property. It follows that these protocols cannot have fully-black-box reductions even if the reductions are tailored and embedding.

---

say that g is uniquely determined if for every input $x \in L$ and every two witnesses w_1, w_2 for x, $g(w_1) = g(w_2)$. (A special case is the function $g(w) = w$ that is uniquely determined in the case where every $x \in L$ has one witness). Our lower bounds apply to any reduction that establishes witness-hiding of some uniquely determined feature of the input, and our results in Section 3 are stated using this terminology.

1.7 Transcript Knowledge Extractors

We now discuss the TKE assumption mentioned earlier. Loosely speaking a reduction for an interactive proof from an hardness assumption (e.g., the existence of bit commitment schemes) has a ***Transcript Knowledge Extractor*** (TKE) if the following holds: There is a polynomial-time machine E that has access to an oracle that breaks commitments, and E is able to extract witnesses from "most" accepting transcripts between any prover P^* and the verifier V. The precise definition is given in Definition 3.8. In many cases (e.g., 3-colorability and Hamiltonicity) the soundness analysis of the protocol implicitly presents a transcript knowledge extractor. More details are given in Section 3.4.

What about non-black-box techniques? Our results only apply when the proof establishing witness-hiding is done by a black-box reduction. As we explained earlier, following the breakthrough paper of Barak [3] there are examples of protocols where the reduction proving zero-knowledge is non-black-box and relies on the code of V^*. We remark that [4] shows that parallel repetition of Hamiltonicity *is* ZK assuming that CIRCUIT-SAT has "small" circuits. Thus, we cannot expect unconditional results that rule out non black-box reductions establishing WH of this protocol. A natural question (that is not addressed in this paper) is to try and prove impossibility results for non-black-box reductions under hardness assumptions.

Organization of this paper. Due to space limitations this extended abstract does not contain all our results and there are no proofs. The reader is referred to the full version for more details. We give formal definitions for WH in Section 2 and our results are stated in Section 3.

2 Definitions of Witness Hiding

2.1 Preliminaries on Interactive Proofs

We use standard definitions of interactive machines and protocols. The reader is referred to [11] for an extensive treatment that also introduces this notation.

In this paper we are only interested in interactive proofs for languages L in NP. Such languages are defined using a *witness relation* R_L (that is L contains all inputs x such that there exist $w \in R_L(x)$). When we consider $L \in$ NP we always assume that it comes with some specific witness relation R_L and that on input $x \in L$ the prover in the interactive proof is provided with some witness $w \in R_L(x)$. We want that completeness, soundness and privacy requirements are maintained for every choice of this witness. We use the following definition.

Definition 2.1 (interactive proofs for NP languages). *Let L be a language in* NP*. A witness choice is a function that maps every input x in L to a random variable W that is distributed over $R_L(x)$. A pair (P, V) of interactive machines is an* interactive proof *for L with completeness $c(n)$ and soundness $s(n)$ if V is probabilistic-polynomial-time and the following two conditions hold:*

– *Completeness: For every $x \in L$ and witness choice W,* $\Pr[(P(W(x)), V)(x) = 1] \geq c(|x|)$.
– *Soundness: For every $x \notin L$ and machine P^*,* $\Pr[(P^*, V)(x) = 1] \leq s(|x|)$.

If the completeness and soundness parameters are omitted then we mean perfect completeness (that is $c(n) = 1$) and negligible soundness (that is $s(n) = neg(n)$).

An interactive proof is public-coin if every message of V consists of independent random coins. The number of rounds in the protocol is the overall number of messages.

2.2 The Concept of Witness Hiding

Witness-hiding interactive proofs (defined by [9]) have the following property: if a verifier can find a witness to the NP statement that is being proven following an interaction with a prover, then he could have done so without such an interaction. This notion is defined with respect to a distribution ensemble over inputs in L.

Definition 2.2. *Let L be a language in* NP*. A distribution ensemble $X = \{X_n\}$ is over* positive instances with respect to *L if for every n, X_n assigns positive probability only to instances in $L \cap \{0,1\}^n$.*

Definition 2.3 (witness-hiding). *Let $L \in$ NP and let R_L be its witness relation. Let $X = \{X_n\}$ be a distribution ensemble over positive instances. An interactive proof (P, V) for L is* witness-hiding *with respect to X if the following condition holds: If for every sufficiently large n and every polynomial size circuit C,*

$$\Pr_{X \leftarrow X_n}[C(X) \in R_L(X)] = neg(n)$$

then for every polynomial-time V^*, every witness choice W, sufficiently large n and every auxiliary input z_n,

$$\Pr_{X \leftarrow X_n}[(P(W(X)), V^*(z_n))(X) \in R_L(X)] = neg(n)$$

Note that there is an inherent difference between the definition of witness-hiding proofs and zero-knowledge proofs in the sense that the definition is with respect to an ensemble X, whereas in zero-knowledge proofs (or witness-indistinguishable proofs) information should not leak on *any* input x.

2.3 Hiding Features of the Witness

Definition 2.3 is only concerned with whether V^* can recover an *entire* witness. A stronger privacy requirement is that V^* does not learn some efficiently computable feature of a witness. Our results are stated using this more general notion. We use the following definition.

Definition 2.4 (feature function). *Let L be a language in* NP *and let $m(n)$ denote the length of witnesses $w \in R_L(x)$ for inputs $x \in L$ of length n. Let $\ell(n)$ be an integer function. A feature function g is a polynomial-time computable function $g : \{0,1\}^{m(n)} \to \{0,1\}^{\ell(n)}$. We say that g is* uniquely determined *on an input $x \in L$, if $w_1, w_2 \in R_L(x)$ implies $g(w_1) = g(w_2)$. In that case, we sometimes abuse the notation and write $g(x)$ rather than $g(w)$. We say that g is* uniquely determined *on a distribution X that is distributed over $L \cap \{0,1\}^n$, if it is uniquely determined on every input in the support of X.*

Using this terminology, we can define the following notion of witness-hiding proofs that hides a uniquely determined feature g of the witness. (We restrict our attention to uniquely determined features as otherwise the feature of a witness $W(X)$ *depends* on the witness choice W). Loosely speaking, the definition below says that if the verifier V^* can distinguish the feature $g(X)$ from uniform following an interaction with the prover then he can do that prior to the interaction.

Definition 2.5 (witness-hiding for a uniquely determined feature g). *Let $L \in$ NP and let R_L be its witness relation. Let $X = \{X_n\}$ be a distribution ensemble over positive instances of L. An interactive proof (P, V) for L is* witness-hiding *a feature g that is uniquely determined with respect to X, if the following condition holds: If for every sufficiently large n and every polynomial size circuit C,*

$$|\Pr_{X \leftarrow X_n}[C(X) = g(X)] - 2^{-\ell(n)}| = neg(n)$$

then for every polynomial-time V^*, every witness choice W, sufficiently large n and every auxiliary input z_n,

$$|\Pr_{X \leftarrow X_n}[(P(W(X)), V^*(z_n))(X) = g(X)] - 2^{\ell(n)}| = neg(n)$$

Remark 2.1 (The case of one witness). In the case that a language $L \in$ NP is defined using a witness relation R_L where every $x \in L$ has exactly one witness then any feature g is uniquely determined. This in particular applies to the feature $g(w) = w$. With this choice definitions 2.5 and 2.3 coincide. Our lower bounds apply to every uniquely determined feature of witnesses and in particular apply to the standard notion of witness-hiding in the case that there is only one witness.

3 Black-Box Witness-hiding and Our Results

We study reductions that establish the conditions of Definitions 2.3 and 2.5. Consider an interactive proof (P, V). A black-box reduction R that establishes the witness-hiding property of (P, V) is a polynomial-time machine that receives oracle access to a "cheating verifier" V^* (that is not necessarily efficient). It is assumed that V^* is able to break the witness-hiding property of the proof system and learn the feature $g(X)$ following an interaction with P. As our goal

is to prove lower bounds on reductions we make it easier for the reduction and assume that V^* learns $g(X)$ with probability one (this only makes our results stronger). The reduction R is required to perform one of the following two tasks when given oracle access to such a V^*:

- Learn the feature $g(X)$ with noticeable advantage when given X as input. (This shows that V^* could have learned $g(X)$ without interacting with the prover).
- Break the security assumption on which the protocol is based. (This gives a contradiction in case V^* is efficient).

We distinguish between two kinds of constructions of interactive protocols depending on whether the protocol relies on a generic security assumption (e.g., "there exist one-way functions" or "there exist bit-commitment schemes") or on a specific security assumption (e.g., "factoring is a one-way function"). This distinction is described in the next sections.

3.1 Weakly Black-box Reductions

In this paper we consider the following notion of weakly-black-box reductions. The "protocol designer" chooses a specific "hardness assumption" (which we model, without lost of generality, as a one-way function) and designs specific machines P, V to be used by the prover and verifier. The designer also chooses a language $L \in \mathrm{NP}$, an ensemble X over positive instances and a feature g that is uniquely determined for X. His goal is to show that (P, V) is witness-hiding for these specific choices and this allows the reduction R to depend in an arbitrary (non black-box) way on all the previous choices. A precise definition follows:

Definition 3.1 (weakly black-box reduction establishing witness hiding). *Let f be a length preserving function. Let L be a language in* NP*. Let (P, V) be a proof system for L. Let $X = \{X_n\}$ be a distribution ensemble over positive instances of L, and g be a feature that is uniquely determined for X. We say that R is a* weakly-black-box WH reduction from *f if R is a polynomial-time oracle machine and there exist polynomials $p(n)$ and $k(n)$ such that for every input length n, and every deterministic (not necessarily efficient) algorithm V^*: If there is a witness choice $W(x)$ such that $\Pr_{X \leftarrow X_n}[(P(W(X)), V^*)(X) = g(X)] = 1$ then*

- *either R^{V^*} inverts f on random inputs of length $k(n)$ with probability $1/p(n)$,*
- *or $\Pr_{X \leftarrow X_n}[R^{V^*}(X) = g(X)] \geq 2^{-\ell(n)} + 1/p(n)$.*

Remark 3.1 (Relationship to black-box simulation). It is natural to compare our definitions to that of "black-box simulation" introduced in [12]. The notion of black-box simulation corresponds to a specific protocol (P, V) and requires that there is one reduction R (called black-box simulator) so that for every *efficient* V^*, $R^{V^*}(x)$ simulates a transcript that is indistinguishable from $(P, V^*)(x)$. It turns out that all black-box simulators in the literature satisfy a stronger requirement: For every V^* (not necessarily efficient) either $R^{V^*}(x)$ simulates a transcript or it is able to invert some one-way function f. Note that every such reduction is a weakly-black-box WH reduction from f.

3.2 Our Results on Weakly Black-box Reductions

We consider several notions of weakly-black-box reductions. A reduction R is *oblivious* if it does not depend on the choice of the distribution ensemble X and one reduction applies to any distribution ensemble. We remark that all proofs of "black-box ZK" [12] in the literature are oblivious reductions.

Definition 3.2 (oblivious reductions). *Let L, P, V, g, f be as in Definition 3.1. Let R be a polynomial-time oracle procedure. We say that R is an* oblivious reduction *if for every distribution ensemble X over positive instances, R is a weakly-black-box* WH *reduction from f with respect to X.*

We show that assuming NP $\neq$ BPP, oblivious reductions cannot show witness-hiding for constant-round public-coin protocols with negligible soundness, with respect to NP complete languages L, where every input $x \in L$ has exactly one witness. This result (stated below) is an easy extension of the negative results of [12] for black-box ZK.

Theorem 3.1. *Let L be a language in* NP *and let R_L be its witness relation. Let (P, V) be a constant-round public-coin interactive proof for L with negligible soundness. Assume that the feature $g(w) = w$ is uniquely determined on every input x (that is that every $x \in L$ has exactly one witness). Let R be an oblivious weakly-black-box* WH *reduction from some one-way function f. Then $L \in$* BPP.

We now consider reductions that can be tailored to a specific distribution ensemble X. Such a weakly-black-box reduction R receives an input x and oracle access to a "cheating verifier" V^* that breaks the witness-hiding property. When R queries V^* it supplies some partial history of the protocol (P, V) and V^* replies with his next message in the protocol. A part of the partial history is the input x' to the protocol (P, V). Note that R may query V^* on partial histories that contain inputs x' that are different from the input x given to R. A reduction R is non-embedding if it finishes all queries to V^* on one input before it queries V^* on some other input.

Definition 3.3 (non-embedding reductions). *Let L, P, V, g, f, X be as in Definition 3.1 and let R be a weakly-black-box* WH *reduction from f. We say that R is a* non-embedding reduction *if for every input x and every oracle V^* and every two inputs $x_1 \neq x_2$, if $R^{V^*}(x)$ makes a query containing x_1 before making a query containing x_2 then all queries that contain x_1 are made before the first query that contains x_2.*

We show that if a non-embedding reduction is used to show witness-hiding for a constant-round public-coin protocol with negligible soundness with respect to some distribution X and uniquely determined feature g, and if furthermore the protocol is also a proof of knowledge, then it is possible to efficiently predict $g(x)$ with noticeable advantage when given x sampled from X. This means that it is impossible to use such reductions to hide features that are hard to predict.

Theorem 3.2. *Let $L \in$ NP, let R_L be its witness relation. Let X be a distribution ensemble over positive instances of L and let g be a feature that is uniquely determined with respect to X. Let (P, V) be a constant-round public-coin interactive proof for L with negligible soundness and assume that (P, V) is a proof of knowledge with negligible knowledge error (see Definition 3.7). Let R be a non-embedding weakly-black-box* WH *reduction from a one-way function f. Then there is a polynomial-time machine M and a polynomial p such that for every sufficiently large n, $\Pr_{X \leftarrow X_n}[M(X) = g(X)] \geq 2^{-\ell(n)} + 1/p(n)$.*

In particular, if every input $x \in L$ has one witness then the feature $g(w) = w$ is uniquely determined. The theorem says that if (P, V) is a proof of knowledge then the existence of a non-embedding reduction R gives that one can efficiently find witnesses when given x sampled from X with noticeable probability (and thus X is not a "hard distribution").

Corollaries on specific protocols. Consider parallel repetition of 3-Colorability [13] and Hamiltonicity [5] using any choice of commitment scheme (that may be based on an arbitrary one-way function). These protocols are constant-round public-coin interactive proofs with negligible soundness for complete languages in NP. Furthermore, both these protocols are proofs of knowledge with negligible knowledge error. Thus, Theorems 3.1 and 3.2 apply and give limitations on reductions that establish the WH property of these protocols.

3.3 Fully-black-box Reductions

In a fully-black-box construction the protocol designer is given a cryptographic primitive as a black-box. (In this paper we consider the primitives: one-way function, one-way permutation and information theoretically binding bit commitments). In this setup the protocol designer receives a black-box that implements the basic primitive. He designs oracle machines $P^{(\cdot)}, V^{(\cdot)}$ to be used by the prover and verifier. We start by formally defining this setup.

Definition 3.4 (black-box interactive proofs). *Let L be language in* NP. *Let $\mathcal{F}$ be a set of functions from strings to strings. A pair $(P^{(\cdot)}, V^{(\cdot)})$ of oracle machines is a $\mathcal{F}$-*black-box interactive proof *for L if V is probabilistic polynomial time and for every $f \in \mathcal{F}$, the pair (P^f, V^f) satisfy the completeness and soundness properties in Definition 2.1.*

We now consider several families of possible oracles that model one-way functions, one-way permutations and bit-commitment schemes. The same framework, however, can be used to describe most cryptographic primitives.

Definition 3.5 (oracles for primitives). *Let O_{OWF} denote the set of all length preserving functions. Let O_{OWP} be the subset of all functions in O_{OWF} that are permutations on every input length. Given $f \in O_{\mathrm{OWF}}$, an algorithm T η-breaks f on security parameter k if $\Pr_{X \leftarrow U_k}[T(f(X)) \in f^{-1}(f(X))] \geq \eta$.*

Let O_{BC} denote the set of all functions f that given a bit b and a string $r \in \{0,1\}^k$ produce a string $c \in \{0,1\}^k$. We furthermore require that f is binding, namely that for every k and $r_1, r_2 \in \{0,1\}^k$, $f(0,r_1) \neq f(1,r_2)$. Given $f \in O_{\mathrm{BC}}$, an algorithm T η-breaks f on security parameter k if $\Pr_{B \leftarrow U_1, R \leftarrow U_k}[T(f(B,R) = B] \geq 1/2 + \eta/2$.[10]

Remark 3.2 (interactive commitment schemes). The family O_{BC} defined above corresponds to perfectly binding non-interactive commitment schemes. In such a scheme the sender commits to a bit b by sending $f(b,r)$ for a randomly chosen r. The sender can later reveal the bit b by sending r and our definition requires that the commitment is binding.

One can consider more relaxed notion of commitment schemes in which the commitment phase is an interactive protocol between the sender and receiver. In such a scheme the binding property can be statistical rather than perfect (namely, binding only holds with high probability over the receiver's coins). We have chosen the more simple version of commitment schemes in order to simplify the presentation. All our results, however, apply also for the more general notion of interactive statistically binding commitments (and this holds by exactly the same proofs).

Remark 3.3 (3-Colorability and Hamiltonicity). Using this framework the classical protocols for 3-Colorability and Hamiltonicity can be viewed as O_{BC}-black-box interactive proofs. (This also applies if we modify O_{BC} to capture interactive commitments as explained in Remark 3.2).

We can now give the definition of a fully-black-box reduction. We consider two flavors depending on whether the black-box interactive proof starts from one-way functions or bit-commitment (that is whether f is assumed to come from O_{OWF} or O_{BC}). The definition below is identical to definition 3.1 with the following modifications: all parties (including the verifier V^* and the reduction R) get oracle access to f and the reduction should work for every choice of f in the family of relevant oracles.

Definition 3.6 (fully-black-box reduction establishing witness hiding). *Let L be a language in* NP. *Let $(P^{(\cdot)}, V^{(\cdot)})$ be a O_{OWF}-black-box interactive proof for L (resp., O_{BC}-black-box interactive proof for L). Let $X = \{X_n\}$ be a distribution ensemble over positive instances of L, and g be a feature that is uniquely determined for X. We say that R is a* fully-black-box WH reduction from OWF *(resp.,* fully-black-box WH reduction from BC*) if R is a polynomial-time oracle machine and there exist polynomials $p(n)$ and $k(n)$ such that for every $f \in O_{\mathrm{OWF}}$ (resp., every $f \in O_{\mathrm{BC}}$) and every input length n, and every deterministic (not necessarily efficient) algorithm V^*: If there is a witness choice $W(x)$ such that* $\Pr_{X \leftarrow X_n}[(P^f(W(X)), V^{*f})(X) = g(X)] = 1$ *then*

[10] The choice of dividing η by 2 is so that the success probability of T is one when $\eta = 1$. This way, for both $O_{\mathrm{OWF}}, O_{\mathrm{BC}}$ an algorithm T that 1-breaks f succeeds with probability one.

– *either* $R^{V^{*f},f}$ $1/p(n)$*-breaks* f *on security parameter* $k(n)$*,*
– *or* $\Pr_{X \leftarrow X_n}[R^{V^{*f},f}(X) = g(X)] \geq 2^{-\ell(n)} + 1/p(n)$*.*

We note that any fully-black-box reduction R gives a weakly-black-box reduction for any specific choice of f.

3.4 Transcript Knowledge Extractors

We introduce a non-standard notion of proofs of knowledge (which is incomparable to the standard one) and show that black-box interactive proofs from commitment schemes that are constant-round public-coin protocols with negligible soundness, and in addition have "transcript knowledge extractors" cannot have fully-black-box reductions establishing WH. Before we define this new notion, let us first recall the definition of "standard" knowledge extractors.

Definition 3.7 (knowledge extractor [11]). *Let* (P, V) *be an interactive proof system for* $L \in$ NP *and let* R_L *be its witness relation. A probabilistic machine* E *is a* knowledge extractor *for* (P, V) *and* R_L *with error* $\eta \colon \mathbb{N} \mapsto \mathbb{R}$*, if there exists a polynomial* q_E *such that for every input* $x \in L_n$ *and every deterministic algorithm* P^**,* $E^{P^*}(x)$ *runs in expected number of step bounded by* $\frac{q_E(n)}{\delta(x) - \eta(|x|)}$ *and outputs* $w \in R_L(x)$*, where* $\delta(x) = \Pr[(P^*, V)(x) = 1]$*.*

The new notion applies to black-box interactive proofs (See definition 3.4) and allow the extractor to access an oracle that breaks the security assumption on which the protocol is based. The extractor gets as input a transcript on which V accepts and is required to extract a witness from the transcript (we stress that the extractor does not get oracle access to the prover P^*). A precise definition follows. The definition has two flavors depending on whether the black-box interactive proofs is from one-way functions or bit-commitment.

Definition 3.8 (transcript knowledge extractor (TKE)). *Let* L *be a language in* NP *and let* $(P^{(\cdot)}, V^{(\cdot)})$ *be a* O_{OWF}*-black-box interactive proof for* L *(resp., a* O_{BC}*-black-box interactive proof for* L*). A polynomial-time oracle machine* E *is a* transcript knowledge extractor *with error* $\eta(n)$ *if for every* $f \in O_{\mathrm{OWF}}$ *(resp., every* $f \in O_{\mathrm{BC}}$*) and every algorithm* T *that* 1*-breaks* f *on every security parameter* k *it holds that: For every input* $x \in L$ *and for every deterministic algorithm* P^**, let* $\tau(x)$ *be the random variable that is the transcript of* $(P^{*f}, V^f)(x)$ *then:*

$$\Pr[\tau(x) \text{ is accepting and } E^{f,T}(\tau(x)) \notin R_L(x)] \leq \eta(|x|)$$

We allow E to access both f and an oracle T that completely breaks f. While transcript knowledge extractors require an oracle that breaks the security assumption, they have the advantage that they do not require oracle access to the prover P^*. This in particular means that they do not rely on rewinding P^* and that the extraction process is efficient even if P^* is inefficient.

3.5 Our Results on Fully-black-box Reductions

We now state our main result on fully-black-box reductions. We consider black-box interactive proofs that use one-way functions or commitment schemes. This result applies to any reduction (even one that is embedding) whenever the black-box interactive proof has a transcript knowledge extractor.

Theorem 3.3. *Let $L \in$ NP and let R_L be its witness relation. Let X be a distribution ensemble over positive instances of L and let g be a feature that is uniquely determined with respect to X. Let $(P^{(\cdot)}, V^{(\cdot)})$ be a constant-round public-coin O_{OWF}-black-box interactive proof for L (resp., O_{BC}-black-box interactive proof for L). Assume that the proof system has negligible soundness and a* TKE *with negligible error. Let R be a fully-black-box* WH *reduction from* OWF *(resp.,* BC*). Then, there is a polynomial-time machine M and a polynomial p such that for every sufficiently large n, $\Pr_{X \leftarrow X_n}[M(X) = g(X)] \geq 2^{-\ell(n)} + 1/p(n)$.*

The theorem above is very similar to Theorem 3.2 with the exception that it handles general fully-black-box reductions (rather than non-embedding weakly-black-box ones) and requires transcript knowledge extractors (rather than standard knowledge extractors). In the next section we observe that many protocols in the literature have transcript knowledge extractors. In particular, when considering a language L in which every $x \in L$ has exactly one witness, the feature $g(w) = w$ is uniquely determined and the Theorem asserts that one cannot use a fully-black-box reduction to establish WH for distributions X for which finding a witness is hard.

3.6 Prevalence of Transcript Knowledge Extractors

On an intuitive level one can expect that any interactive proof where the privacy of the prover is based on a hardness assumption (e.g., the existence of bit-commitment schemes) has a transcript knowledge extractor as otherwise the hardness assumption is "not really needed" and the security of the protocol follows unconditionally. We do not make such a formal statement and do not know whether a statement of this flavor is true. In the discussion below we observe that many specific interactive proofs in the literature have transcript knowledge extractors. The impossibility results of Theorem 3.3 apply to all these protocols.

3-Colorability. Consider the ZK proof of [13] for 3-colorability. This is a O_{BC}-black-box interactive proof that is a 3-round protocol with perfect completeness and soundness $1 - 1/m$ (where m is the number of edges in the input graph). It is known that this protocol is zero-knowledge. The soundness analysis of this protocol shows that if in the first message of the protocol the prover does not send a commitment to a witness (a legal coloring) then with probability $1/m$ (where m is the number of edges in the input graph) the verifier rejects. It follows that this protocol has a transcript knowledge extractor with $\eta = 1 - 1/m$ as E can open the commitment using the fact it has oracle access to an algorithm T

that breaks the commitment. Recall that we are interested in investigating the security of the parallel repetition of this atomic protocol when repeated t times. It is easy to see that after repetition there is a transcript knowledge extractor with error $\eta = (1 - 1/m)^t$. (This follows as if the extractor cannot find a witness in any of the commitments sent in the first round then the probability that the verifier accepts is the expression above).

Graph Hamiltonicity. Consider the ZK proof of [5] for Graph Hamiltonicity. This is a O_{BC}-black-box interactive proof that is a 3-round protocol with perfect completeness and soundness $1/2$. The soundness analysis of this protocol shows that if B does not commit to a graph that is isomorphic the input graph G in its first message then with probability $1/2$ he is caught in the third message. On the other hand if B commits to a graph that is isomorphic to the correct graph then with probability half he reveals a cycle in the graph in the third message and the knowledge extractor can "break" the commitment and find a cycle in the original graph when given the transcript. These properties give a knowledge extractor with $\eta = 1/2$ and parallel repetition reduces η at an exponential rate.

Zaps. These are 2-message WI protocols [7], and are not known to be a (standard) proof of knowledge. Zaps can be either constructed based on non-interactive zero-knowledge (NIZK) proofs, or based on a verifiable pseudo-random generator (VPRG). The "generic" versions of both of these primitives are constructed using trapdoor permutations, where the role of trapdoor permutations in all known constructions is to implement the *hidden bits* (or *hidden random string*) model [8, 16, 7]. A close examination of these constructions reveals that if one is able to invert the underlying trapdoor permutation then the bits (random string) becomes completely revealed. As observed in [19], such information can be used to extract the witness for the statement. With appropriately chosen parameters (i.e., if the soundness error is small enough), this can be done with all but negligible probability. The same applies for VPRG based constructions. Thus, many of the "generic" zap constructions have transcript knowledge extractors.

Acknowledgements We thank Oded Goldreich and Rafael Pass for helpful discussions.

References

1. A. Akavia, O. Goldreich, S. Goldwasser, and D. Moshkovitz. On basing one-way functions on np-hardness. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC)*, pages 701–710, 2006.
2. L. Babai and S. Moran. Arthur-merlin games: A randomized proof system, and a hierarchy of complexity classes. *J. Comput. Syst. Sci.*, 36(2):254–276, 1988.
3. B. Barak. How to go beyond the black-box simulation barrier. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–115, 2001.

4. B. Barak, Y. Lindell, and S. Vadhan. Lower bounds for non-black-box zero knowledge. *Journal of Computer and System Sciences*, 72(2):321–391, 2006.
5. M. Blum. How to prove a theorem so no one else can claim it. In *Proceedings of the International Congress of Mathematicians*, pages 1444–1451, 1987.
6. A. Bogdanov and L. Trevisan. On worst-case to average-case reductions for np problems. *SIAM Journal on Computing*, 36(4):1119–1159, 2006.
7. C. Dwork and M. Naor. Zaps and their applications. *SIAM Journal on Computing*, 36(6):1513–1543, 2007.
8. U. Feige, D. Lapidot, and A. Shamir. Multiple noninteractive zero knowledge proofs under general assumptions. *SIAM Journal on Computing*, 29(1):1–28, 1999.
9. U. Feige and A. Shamir. Witness indistinguishable and witness hiding protocols. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing (STOC)*, pages 416–426. ACM, 1990.
10. J. Feigenbaum and L. Fortnow. Random-self-reducibility of complete sets. *SIAM Journal on Computing*, 22(5):994–1005, 1993.
11. O. Goldreich. *Foundations of Cryptography: Basic Tools*. Cambridge University Press, 2001.
12. O. Goldreich and H. Krawczyk. On the composition of zero-knowledge proof systems. *SIAM J. Comput.*, 25(1):169–192, 1996. Preliminary version in *ICALP'90*.
13. O. Goldreich, S. Micali, and A. Wigderson. Proofs that yield nothing but their validity and a methodology of cryptographic protocol design (extended abstract). In *FOCS*, pages 174–187. IEEE, 1986.
14. O. Goldreich and Y. Oren. Definitions and properties of zero-knowledge proof systems. *Journal of Cryptology*, 7(1):1–32, 1994.
15. S. Goldwasser, S. Micali, and C. Rackoff. The knowledge complexity of interactive proof systems. *SIAM J. Comput.*, 18(1):186–208, 1989. Preliminary version in *STOC'85*.
16. J. Kilian and E. Petrank. An efficient noninteractive zero-knowledge proof system for np with general assumptions. *J. Cryptology*, 11(1):1–27, 1998.
17. R. Pass. Parallel repetition of zero-knowledge proofs and the possibility of basing cryptography on np-hardness. In *IEEE Conference on Computational Complexity*, pages 96–110, 2006.
18. O. Reingold, L. Trevisan, and S. P. Vadhan. Notions of reducibility between cryptographic primitives. In *Theory of Cryptography, First Theory of Cryptography Conference, TCC 2004*, volume 2951 of *Lecture Notes in Computer Science*, pages 1–20. Springer, 2004.
19. A. D. Santis and G. Persiano. Zero-knowledge proofs of knowledge without interaction. In *Proceedings of the 33rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 427–436, 1992.