Cryptographic reverse firewalls for interactive proof systems

DOI:
10.1016/j.tcs.2020.11.043

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Published in:
Theoretical Computer Science

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6B0] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Cryptographic Reverse Firewalls for Interactive Proof Systems

Chaya Ganesh  
Department of Computer Science and Automation, Indian Institute of Science, India  
chaya@iisc.ac.in

Bernardo Magri  
Department of Computer Science, Aarhus University, Denmark  
magri@cs.au.dk

Daniele Venturi  
Department of Computer Science, Sapienza University of Rome, Italy  
venturi@di.uniroma1.it

Abstract
We study interactive proof systems (IPSes) in a strong adversarial setting where the machines of honest parties might be corrupted and under control of the adversary. Our aim is to answer the following, seemingly paradoxical, questions:

Can Peggy convince Vic of the veracity of an NP statement, without leaking any information about the witness even in case Vic is malicious and Peggy does not trust her computer?

Can we avoid that Peggy fools Vic into accepting false statements, even if Peggy is malicious and Vic does not trust her computer?

At EUROCRYPT 2015, Mironov and Stephens-Davidowitz introduced cryptographic reverse firewalls (RFs) as an attractive approach to tackling such questions. Intuitively, a RF for Peggy/Vic is an external party that sits between Peggy/Vic and the outside world and whose scope is to sanitize Peggy’s/Vic’s incoming and outgoing messages in the face of subversion of her/his computer, e.g. in order to destroy subliminal channels.

In this paper, we put forward several natural security properties for RFs in the concrete setting of IPSes. As our main contribution, we construct efficient RFs for different IPSes derived from a large class of Sigma protocols that we call malleable.

A nice feature of our design is that it is completely transparent, in the sense that our RFs can be directly applied to already deployed IPSes, without the need to re-implement them.

2012 ACM Subject Classification  Security and privacy → Cryptography

Keywords and phrases  Subversion, Algorithm substitution attacks, Cryptographic reverse firewalls, Interactive proofs, Zero knowledge

Digital Object Identifier  10.4230/LIPIcs.ICALP.2020.55

Related Version  A full version of the paper is available at https://eprint.iacr.org/2020/204.

Funding  Bernardo Magri: This work was supported by Concordium Blockchain Research Center, Aarhus University, Denmark
1 Introduction

An interactive proof system (IPS) $\Pi = (P, V)$ allows a prover $P$ to convince a verifier $V$ about the veracity of a public statement $x \in L$, where $L$ is an NP language and where both $P$ and $V$ are modeled as interactive PPT machines. The prover is facilitated by possessing a witness $w$ to the fact that, indeed, $x \in L$, and the interaction with the verifier may consist of several rounds of communication, at the end of which the verifier outputs a verdict on the membership of $x$ in $L$.

In order to be useful, an IPS should satisfy the following properties:

- **Completeness:** If $x \in L$, the honest prover (almost) always convinces the honest verifier.
- **Soundness:** If $x \notin L$, no (computationally bounded) malicious prover can convince the honest verifier that $x \in L$. An even stronger guarantee, known as knowledge soundness [9], is to require that the only way a prover can convince the honest verifier that $x \in L$ is to “know” a valid witness $w$ corresponding to $x$. Such proofs\(^1\) are called proofs of knowledge (PoKs).
- **Zero Knowledge (ZK):** A valid proof reveals nothing beyond the fact that $x \in L$, and thus in particular it leaks no information about the witness $w$, even in case the proof is conducted in the presence of a (computationally bounded) malicious verifier [36]. A weaker guarantee, known as witness indistinguishability (WI) [24], is that, whenever there are multiple witnesses attesting that $x \in L$, no (computationally bounded) malicious verifier can distinguish whether a proof is conducted using either of two witnesses.

One of the motivations for studying IPSes with the above properties is that they are ubiquitous in cryptography, with applications ranging from identification protocols [24], blind digital signatures [42], and electronic voting [16], to general-purpose maliciously secure multi-party computation [35].

1.1 Sigma Protocols

While WI/ZK PoKs exist for all of NP, based on minimal cryptographic assumptions [23, 34, 33], efficiency is a different story. Fortunately, it is possible to design practical interactive proofs for specific languages, typically in the form of so-called Sigma protocols. Briefly, a Sigma protocol is a special type of IPS consisting of just three rounds, where the prover sends a first message $\alpha$ (the commitment), the verifier sends a random string $\beta$ (the challenge), and finally the prover forwards a last message $\gamma$ (the response). Sigma protocols satisfy two main properties: The first one, known as special soundness, is a strong form of knowledge soundness; the second one, known as honest-verifier zero knowledge (HVZK), is a weak form of the zero knowledge property that only holds against honest-but-curious verifiers.

The applications of Sigma protocols to cryptographic constructions are countless (see, e.g., [25, 17, 48, 22, 43]). These results are perhaps surprising, as Sigma protocols only satisfy HVZK and thus guarantee no security in the presence of malicious verifiers. In some cases, the solution to this apparent paradox is due to a beautiful technique put forward by Cramer, Damgård, and Schoenmakers [15], which allows to add WI to any Sigma protocol. Moreover, it is relatively easy to transform any Sigma protocol into an interactive ZK PoK at the cost of adding a single round of interaction [33].

\(^1\) Sometimes, the term “proof” is used to refer to statistically sound IPSes, while computationally sound IPSes are typically called “arguments”.
1.2 Our Question

The standard definitions of security for IPSes (implicitly) rely on the assumption that honest parties can fully trust their machines. In practice, however, such an assumption may just be too optimistic, as witnessed by the revelations of Edward Snowden about subversion of cryptographic standards [45, 7], and in light of the numerous (seemingly accidental) bugs in widespread pieces of cryptographic software [38, 1, 2].

Motivated by the above incidents, we ask the following question which constitutes the main source of inspiration for this work:

Can we design practical interactive proofs that remain secure even if the machines of the honest parties running them have been tampered with?

In order to see why the above question is well motivated and not trivial, let us analyze the dramatic consequences of subverting the prover of ZK IPSes. Clearly, the problem of subversion-resistant interactive zero knowledge is just impossible in its utmost generality, as a subverted prover could just reveal the witness to the verifier. However, one may argue that these kind of attacks are easily detectable, and thus can be avoided.

The problem becomes more interesting if we restrict the subversion to be undetectable, as suggested by Bellare, Paterson, and Rogaway [11] in their seminal work on subversion of symmetric encryption, where the authors show how to subvert any sufficiently randomized cipher in an undetectable manner, using rejection sampling. A moment of reflection shows that their attack can be adapted to the case of IPSes. The solution proposed by [11] is to rely on deterministic symmetric encryption. Unfortunately, this approach is not viable for the case of IPSes, as it is well-known that interactive proofs with deterministic provers can be zero knowledge only for trivial languages [32, §4.5].

Reverse firewalls

The above described undetectable attacks show that the problem of designing IPSes that remain secure even when run on untrusted machines is simply impossible if we are not willing to make any further assumption. In this paper, we study how to tackle subversion attacks against interactive proofs in the framework of “cryptographic reverse firewalls (RFs)”, introduced by Mironov and Stephens-Davidowitz [40]. In such a setting, both the prover and the verifier are equipped with their own RF $W$, also modeled as an interactive PPT machine, whose scope is solely to sanitize the parties’ incoming and outgoing messages in the face of subversion.

Importantly, neither the prover nor the verifier put any trust in the RF, meaning that they are not allowed to share secrets with the firewall itself. The hope is that an uncorrupted RF can provide meaningful security guarantees even in case the honest prover’s and/or verifier’s machines have been tampered with. Note that a RF can never “create security”, as it does not even know the inputs to the protocol, but at best can preserve the security guarantees satisfied by the initial IPS. At the same time, the RF should not ruin the functionality of the protocol.

---

2 In particular, a subverted prover with a hardwired secret key $k$ for a pseudorandom function $F_k(\cdot)$, could sample the random coins $r^{(i)}$ needed to generate the honest prover’s message $m^{(i)}$ (for round $i \in \mathbb{N}$) multiple times, until $F_k(m^{(i)})$ leaks one bit of the witness. This attack works provided that at least one of the prover’s messages has high-enough min-entropy.

3 Clearly, if both the machine of the honest party and the firewall are corrupted, there is no hope for security. On the other hand, in case the machine is honest and the firewall is corrupt, the underlying protocol is still secure, since we can simply think of the RF as being part of the adversary [21].
underlying IPS, in the sense that the sanitized IPS should still work in case no subversion takes place.

Mironov and Stephens-Davidowitz construct general-purpose RFs that can be used in order to preserve both functionality and security of any two-party protocol. It is important to note that since ZK/WI IPSes are a special case of secure two-party computation, their RF constructions already seem to solve our problem. However, the solutions in [40] are not practical. In particular, one of their RFs increases the round complexity of the initial IPS, and, more importantly, it requires to carry out the underlying IPS in the encrypted domain, thus requiring to completely change the original protocol. In contrast, we seek constructions of RFs that can be applied directly to existing IPSes, without adding any overhead, and without the need to re-implement them.

2 Reverse Firewalls for Interactive Proofs

In this section, we give security definitions for RFs applied to IPSes. Our notions can be seen as special cases of the generic framework by Mironov and Stephens-Davidowitz [40], who defined security of RFs for the more general case of arbitrary two-party protocols.

Let $\Pi = (P, V)$ be an IPS for a relation $R$. A cryptographic reverse firewall is an external party $W$ that can be attached either to the prover $P$ or to the verifier $V$, whose scope is to sanitize incoming and outgoing messages in the face of parties’ subversion. Importantly, the RF is allowed to keep its own state but cannot share state with any of the parties. Similarly to [40], we model an interactive Turing machine $M = (M_{\text{nxt}}, M_{\text{rec}}, M_{\text{out}})$ specified as follows: (i) Algorithm $M_{\text{nxt}}$ takes as input the current state and outputs the next message to be sent; (ii) Algorithm $M_{\text{rec}}$ takes as input an incoming message, and updates the state; (iii) Algorithm $M_{\text{out}}$ takes as input the final state at the completion of the protocol, and returns a bit.

Definition 1 (RF for IPSes). Let $\Pi = (P, V)$ be an IPS for a relation $R$. A cryptographic reverse firewall (RF) for $\Pi$ is a stateful algorithm $W$ that takes as input a message, its state, and outputs a sanitized message, together with an updated state. For an interactive Turing machine $M = (M_{\text{nxt}}, M_{\text{rec}}, M_{\text{out}}) \in \{P, V\}$, and RF $W$, the sanitized machine $W \circ M := \hat{M} = (\hat{M}_{\text{nxt}}, \hat{M}_{\text{rec}}, \hat{M}_{\text{out}})$ is specified as follows:

\[
\hat{M}_{\text{nxt}}(\sigma) := W(M_{\text{nxt}}(\sigma))
\]

\[
\hat{M}_{\text{rec}}(\sigma, m) := M_{\text{rec}}(\sigma, W(m))
\]

\[
\hat{M}_{\text{out}}(\sigma) := M_{\text{out}}(\sigma).
\]

As our first contribution, we put forward several natural properties that a RF for an IPS might satisfy. In particular, we consider the following notions (see the full version [29] for more formal definitions).

Completeness preservation: The sanitized IPS (i.e., the IPS obtained by sanitizing both the honest prover’s and the honest verifier’s messages) still satisfies completeness.

Strong soundness preservation: Whenever $x \notin L$, no malicious prover can convince the verifier that $x \in L$, even if the verifier’s implementation has been arbitrarily subverted.

\[\text{At least to some extent, since, strictly speaking, their results for IPSes are incomparable to ours. We refer the reader to §5.1 for more details.}\]
Strong ZK preservation: A valid proof reveals nothing beyond the fact that \( x \in \mathcal{L} \), even in case the proof is conducted in the presence of a malicious verifier talking to a prover whose implementation has been arbitrarily subverted.

Strong WI preservation: Whenever there are multiple witnesses attesting that \( x \in \mathcal{L} \), no malicious verifier talking to a prover whose implementation has been arbitrarily subverted can distinguish whether a proof is conducted using either of two witnesses.

Strong exfiltration resistance for the prover (resp. verifier): Transcripts produced by running the sanitized IPS in the presence of a malicious verifier (resp. prover) talking to a prover (resp. verifier) whose implementation has been arbitrarily subverted are indistinguishable to transcripts produced by running the sanitized IPS in the presence of a malicious verifier (resp. prover) talking to the honest prover (resp. verifier).

For each of the above properties (except for completeness), we also consider a weak variant which only holds w.r.t. functionality-maintaining provers/verifiers. Intuitively, a prover is functionality maintaining if, upon input a valid statement/witness pair, it still convinces the honest verifier with overwhelming probability. Similarly, a verifier is functionality maintaining if, upon input a valid statement, it still accepts with overwhelming probability in a protocol run with the honest prover.

What is possible and what is impossible

A moment of reflection shows that soundness preservation is impossible to achieve. In fact, an arbitrarily subverted verifier might always\(^5\) output one, thus automatically accepting both true and false statements. Such a verifier is still functionality maintaining,\(^6\) and thus this simple attack even rules out weak soundness preservation. One way to circumvent this impossibility would be to only consider partial subversion, i.e. split the verifier into two components, one for computing the next messages in the protocol, and the other one for determining the final verdict on the veracity of a statement; hence, assume the latter component to be untamperable.

Turning to subversion of the prover, consider the subverted prover that always outputs the all-zero string. The soundness property of the underlying IPS implies that, for any RF and for any false statement \( x \notin \mathcal{L} \), a sanitized transcript in this case can never be accepting. Moreover, assuming the language \( \mathcal{L} \) is non-trivial, the latter holds true even in case \( x \) is a true statement, which in turn rules out strong exfiltration resistance. For similar reasons, strong ZK/WI preservation are also impossible to achieve.

3 Firewall Constructions from Malleable Sigma Protocols

As our second contribution, we formalize a class of Sigma protocols which admit simple, and very efficient, RFs for the prover. (See the full version \([29]\) for similar constructions dealing with functionality-maintaining subversion of the verifier.) The main idea is to use the RF to re-randomize the prover’s messages, in order to destroy any potential subliminal channel signaling information about the witness. The difficulty, though, is that such re-randomization must be carried out without knowing a witness, and while at the same time preserving the completeness property of the underlying IPS.

---

\(^5\) If one insists on undetectability, the subverted verifier may output 1 upon some hard-wired, randomly chosen, false statement \( \mathcal{T} \notin \mathcal{L} \).

\(^6\) The latter is because completeness is a guarantee that only concerns true statements.
Figure 1 Cryptographic reverse firewall for a malleable Sigma protocol

For the sake of concreteness, let us describe our firewall applied to the classical Sigma protocol for proving knowledge of a discrete logarithm [49]. Here, the statement consists of a description of a cyclic group \( G \) with generator \( g \) and prime order \( q \), together with a value \( x \in G \) such that \( x = g^w \) for some \( w \in \mathbb{Z}_q \). The prover’s first message is a random group element \( \alpha = g^a \in G \). Finally, the prover’s last message is \( \gamma = a - w \cdot \beta \), where \( \beta \in \mathbb{Z}_q \) is the verifier’s challenge; the verifier accepts \((\alpha, \beta, \gamma)\) if and only if \( g^\gamma = \alpha \cdot x^{-\beta} \). Our RF sanitizes the messages \( \alpha \) and \( \gamma \) from a possibly subverted implementation of the prover as follows:

\[
\hat{\alpha} = \alpha \cdot g^\sigma, \\
\hat{\gamma} = \gamma + \sigma,
\]

for random \( \sigma \in \mathbb{Z}_q \). Note that \( g^{\hat{\gamma}} = g^\alpha \cdot g^\sigma \cdot x^{-\beta} = \hat{\alpha} \cdot x^{-\beta} \), and thus the RF preserves completeness.

We now sketch the proof of weak HVZK preservation. Observe that for any \( \hat{\alpha} = g^\alpha \) sent by a functionality-maintaining subverted prover, the distribution of \( \hat{\alpha} = g^{\hat{\alpha} + \sigma} \) is uniform over \( G \) and independent of \( \hat{\alpha}, \hat{\alpha} \), and in fact it is identical to the distribution of \( \alpha \) in an honest run of the original Sigma protocol (without the firewall). As for \( \hat{\gamma} \), note that if there would be two possible values \( \gamma, \gamma' \) which make both \( \tau = (\alpha, \beta, \gamma) \) and \( \tau' = (\alpha, \beta, \gamma') \) valid transcripts, the choice of which response to pick could be used by a functionality-maintaining subverted prover as a subliminal channel signaling information about the witness. Hence, we exploit the fact that for any prefix \( \alpha, \beta \), there exists a unique response \( \gamma \) such that the verifier accepts upon input \( x \) and \( (\alpha, \beta, \gamma) \).

It follows that the distribution of \( \hat{\gamma} \) is identical to that of \( \gamma \) in an honest run of the original Sigma protocol (without the firewall). Putting it all together, we have shown that the distribution of a sanitized transcript \( \hat{\tau} = (\hat{\alpha}, \beta, \hat{\gamma}) \) is identical to the distribution of an honest transcript \( \tau = (\alpha, \beta, \gamma) \). Thus, weak HVZK preservation follows by the fact that Schnorr’s Sigma protocol is HVZK.

### 3.1 HVZK Preservation

Let us now explain how to generalize the above idea to a large class of Sigma protocols that we call *malleable*. In what follows, given a Sigma protocol \( \Sigma = (P, V) \), we denote by \( P_1 \) and \( P_2 \) the algorithms that compute, respectively, the first prover’s message \( \alpha \), and the
As Sigma protocols are not in general zero knowledge, there is no hope to prove that the
prover’s first messages; we assume that membership in \( A \) can be tested efficiently, so that \( V \)
always outputs \( \bot \) whenever \( \alpha \notin A \).

As for the case of Schnorr’s Sigma protocol, an additional requirement that we need is
that the prover’s responses are unique, meaning that for all \( x \in \mathcal{L} \), and for any \( \alpha \in A \) and
\( \beta \in \{0,1\}^\ell \), there exists at most one\(^8\) value \( \gamma \) such that \( V(x,(\alpha,\beta,\gamma)) = 1 \).

Intuitively, a Sigma protocol is malleable if there exists an efficient algorithm Maul for
randomizing the prover’s first message \( \alpha \) into a value \( \hat{\alpha} \) which is distributed identically to
the first message of an honest prover. Moreover, for any challenge \( \beta \), given the coins used
to randomize \( \alpha \) and any response \( \gamma \) yielding a valid transcript \( \tau = (\alpha,\beta,\gamma) \), there exists an
efficient algorithm Bal for computing a balanced response \( \hat{\gamma} \) such that \( (\hat{\alpha},\beta,\hat{\gamma}) \) is also valid.
As we show in the full version \[29\], many natural Sigma protocols are already malleable.
In particular, the latter holds true for Maurer’s unifying protocol \[39\], which includes the
protocols by Fiat-Shamir \[25\], Guillou-Quisquater \[37\], Schnorr \[49\], Okamoto \[41\], and many
others as special cases.

Our RF construction is depicted in Fig. 1. Intuitively, the firewall uses the malleability
property of the underlying Sigma protocol in order to re-randomize the prover’s first and
last messages, in such a way that a functionality-maintaining subverted prover cannot signal
information about the witness through them. The theorem below, whose proof appears in
the full version \[29\], establishes its security.

\[\textbf{Theorem 2.} \] Let \( \Sigma = (P = (P_1, P_2), V) \) be a malleable Sigma protocol with unique responses,
for a relation \( \mathcal{R} \). The RF \( W \) of Fig. 1 preserves completeness, and is weakly HVZK preserving
for the prover.

### 3.2 ZK Preservation

As Sigma protocols are not in general zero knowledge, there is no hope to prove that the
above firewall weakly preserves ZK. However, a standard trick \[33\] allows to transform any
Sigma protocol into a 5-round IPS satisfying ZK. The idea is to let the prover send the public
key \( pk \) of a commitment scheme \( \langle \text{Gen, Com, Open} \rangle \) during the first round. Then, during the
second round, the verifier forwards to the prover a commitment \( c \) to the challenge \( \beta \). Finally,
the Sigma protocol is executed as before with the difference that the verifier also needs to
open the commitment, with the prover aborting if the opening is invalid. We depict such a
modified protocol in Fig. 2.

In order to build a RF for this IPS, we need to sanitize the additional messages from the
(possibly subverted, but functionality-maintaining) prover.\(^9\) We do so by relying on a special
type of \textit{key-malleable} commitment, which intuitively allows to maul any public key \( pk \) (via
an algorithm MaulKey) into a uniformly random public key \( \hat{pk} \), in such a way that, given
a commitment \( c \) with opening \( d \) w.r.t. \( pk \), it is possible to map \( (c,d) \) into a commitment
\( \hat{c} \) with opening \( \hat{d} \) w.r.t. \( \hat{pk} \), without changing the message inside the commitment. We

\[\text{\textsuperscript{7}}\text{In the case of Schnorr’s Sigma protocol, the challenge space is a cyclic group. However, we can embed}
\text{such group in } \{0,1\}^\ell \text{ for some } \ell \in \mathbb{N}.
\text{\textsuperscript{8}}\text{This property is met by many natural Sigma protocols, and was already considered in several previous}
\text{works }[26,22,51].
\text{\textsuperscript{9}}\text{The other messages are sanitized as before, i.e. we still exploit the fact that the underlying Sigma}
\text{protocol is malleable.}
\begin{figure}[h]
\centering
\begin{tabular}{ll}
<table>
<thead>
<tr>
<th>Prover ((x, w))</th>
<th>Verifier ((x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pk \leftarrow \text{Gen}(1^\lambda))</td>
<td>(\beta \leftarrow {0, 1}^\ell)</td>
</tr>
<tr>
<td>((c, d) \leftarrow \text{Com}(pk, \beta))</td>
<td></td>
</tr>
<tr>
<td>(\alpha = P_1(x, w; a))</td>
<td>(\alpha)</td>
</tr>
<tr>
<td>(\beta = \text{Open}(pk, c, d))</td>
<td>(d)</td>
</tr>
<tr>
<td>If (\beta \neq \bot), then (\gamma = P_2(x, w, \alpha, \beta, a))</td>
<td></td>
</tr>
<tr>
<td>Else, (\gamma = \bot)</td>
<td></td>
</tr>
<tr>
<td>(\gamma)</td>
<td></td>
</tr>
<tr>
<td>(V(x, (\alpha, \beta, \gamma)) = 1)</td>
<td></td>
</tr>
</tbody>
</table>
\end{tabular}
\caption{Sigma protocol compiled with standard techniques to obtain full zero knowledge}
\end{figure}

denote by \texttt{MaulCom} and \texttt{BalOpen}, respectively, the algorithms for mauling the commitment \(c\) and the opening \(d\), and additionally require that the distribution of mauled public keys and commitments is identical, respectively, to that of honestly computed public keys and commitments. As we show in the full version [29], the standard Pedersen’s commitment [44] is easily seen to be key malleable, thus yielding a concrete instantiation under the Discrete Logarithm assumption.

Our RF for the protocol of Fig. 2 is depicted in Fig. 3. The theorem below, whose proof appears in the full version [29], establishes its security.

\textbf{Theorem 3.} Let \(\Sigma = (P = (P_1, P_2), V)\) be a malleable Sigma protocol with unique responses, for a relation \(R\). Let \(\Gamma = (\text{Gen}, \text{Com}, \text{Open})\) be a key-malleable commitment scheme with message space \(\{0, 1\}^\ell\). The RF \(W\) of Fig. 3 preserves completeness, and moreover is weakly exfiltration resistant and weakly zero-knowledge preserving for the prover.

\textbf{Remark 4 (On knowledge soundness).} The IPS of Fig. 2 satisfies soundness, but is not in general a proof of knowledge. However, we would like to note that the prover’s firewall still works for the standard transformation of a Sigma protocol into a zero-knowledge proof of knowledge. In such a transformation, a \textit{trapdoor} commitment scheme is used to commit to the verifier’s challenge. Then, after the verifier decommits, the prover sends the trapdoor to the verifier. This allows an extractor to learn the trapdoor, rewind the prover, and open the commitment to a different challenge, thus learning the response for two different challenges, which allows it to obtain a witness using special soundness.

The prover’s RF for this protocol stays the same, except that it additionally needs to provide a trapdoor for the mauled public key \(\hat{pk}\) given a trapdoor for the original public key \(pk\). This is possible, for instance, using Pedersen’s commitment, where given a public key \(pk = (g, h = g^k)\) with trapdoor \(k\), we can maul the key to \((\hat{g} = g^{t_1}, \hat{h} = h^{t_2})\) for random \(t_1, t_2\). Given the trapdoor \(k\) for the key \(pk\), the trapdoor for the mauled key \(\hat{pk}\) can be computed as \(t_2t_1^{-1}k\).


4 Firewalls for Proving Compound Statements

In this section, we show how to construct firewalls for Sigma protocols that prove compound statements.

Given two Sigma protocols $\Sigma_0$ and $\Sigma_1$ for NP languages $L_0$ and $L_1$, it is easy to obtain a Sigma protocol $\Sigma_{\text{AND}}$ for the NP language $L_{\text{AND}} = \{(x_0, x_1) : x_0 \in L_0 \land x_1 \in L_1\}$ by simply running $\Sigma_0$ and $\Sigma_1$ in parallel, with the verifier sending a single challenge. In a similar vein, the OR technique by Cramer, Damgård, and Schoenmakers [15] allows to obtain a Sigma protocol $\Sigma_{\text{OR}}$ for the NP language $L_{\text{OR}} = \{(x_0, x_1) : x_0 \in L_0 \lor x_1 \in L_1\}$. Importantly, if $\Sigma_0$ and $\Sigma_1$ are both perfect HVZK, $\Sigma_{\text{OR}}$ satisfies perfect WI. On the other hand, Garay et al. [30] showed that if $\Sigma_0$ and $\Sigma_1$ are computational HVZK, $\Sigma_{\text{OR}}$ satisfies computational WI, although the latter holds only in case both statements $x_0, x_1$ in the definition of language $L_{\text{OR}}$ are true (but the prover knows either a witness for $x_0$ or for $x_1$).

As long as $\Sigma_0$ and $\Sigma_1$ are malleable, it is easy to build RFs for $\Sigma_{\text{AND}}$ and $\Sigma_{\text{OR}}$ using our techniques. The RF for $\Sigma_{\text{AND}}$ weakly preserves HVZK, whereas the RF for $\Sigma_{\text{OR}}$ weakly preserves both HVZK and WI.

4.1 AND Composition

Given $x_0, x_1$, a prover wishes to prove to a verifier that $x_0 \in L_0$ and $x_1 \in L_1$. More precisely, consider the derived relation:

$$R_{\text{AND}} = \{((x_0, x_1), (w_0, w_1)) : (x_0, w_0) \in R_0 \land (x_1, w_1) \in R_1\}.$$ 

Let $\Sigma_0 = (\Pi_0^1, \Pi_0^2, V^0)$ (resp. $\Sigma_1 = (\Pi_1^1, \Pi_1^2, V^1)$) be a Sigma protocol for language $L_0$ (resp. $L_1$). A Sigma protocol $\Sigma_{\text{AND}}$ for the relation $R_{\text{AND}}$ can be obtained by running the two provers of $\Sigma_0$ and $\Sigma_1$ in parallel, with the verifier sending a single challenge for both
executions. Fig. 4 shows a RF for the prover of $\Sigma_{\text{AND}}$, assuming that both $\Sigma_0$ and $\Sigma_1$ are malleable. We prove the following result, whose proof appears in the full version [29].

**Theorem 5.** Let $\Sigma_0 = (P^0 = (P^0_0, P^0_1), V^0)$ and $\Sigma_1 = (P^1 = (P^1_0, P^1_1), V^1)$ be malleable Sigma protocols with unique responses, for relations $R_0$ and $R_1$. The RF $W$ of Fig. 4 preserves completeness, and is weakly HVZK preserving for the prover of the Sigma protocol $\Sigma_{\text{AND}}$ for relation $R_{\text{AND}}$.

### 4.2 OR Composition

Given $x_0, x_1$, a prover wishes to prove to a verifier that either $x_0 \in L_0$ or $x_1 \in L_1$ (without revealing which one is the case). More precisely, consider the derived relation

$$R_{\text{OR}} = \{(x_0, x_1), (w_0, w_1) : (x_0, w_0) \in R_0 \lor (x_1, w_1) \in R_1\}.$$  

Let $\Sigma_0 = ((P^0_0, P^0_1), V^0)$ (resp. $\Sigma_1 = ((P^1_0, P^1_1), V^1)$) be a Sigma protocol for language $L_0$ (resp. $L_1$); we denote by $S^0$ (resp. $S^1$) the HVZK simulator for $\Sigma_0$ (resp. $\Sigma_1$). A Sigma protocol $\Sigma_{\text{OR}}$ for the relation $R_{\text{OR}}$ has been constructed for the first time in [15], where the authors showed that $\Sigma_{\text{OR}}$ satisfies both (perfect) special HVZK and (perfect) WI. We describe the protocol $\Sigma_{\text{OR}}$ in Fig. 5, and depict our RF for the prover in Fig. 6.

As in the case of AND composition, we still rely on the fact that the input Sigma protocols $\Sigma_0, \Sigma_1$ are malleable. An additional difficulty, however, stems from the fact that a functionality maintaining prover could now try to change the distribution of the challenges $\beta_0, \beta_1$ in such a way that, even if $\beta_0 \oplus \beta_1 = \beta$, the pair $(\beta_0, \beta_1)$ signals some information about the witness $w$ or about the hidden bit $b$. Intuitively, the RF in Fig. 6 tackles this attack by randomizing the challenges $\beta, \beta_0, \beta_1$. The latter requires a different form of malleability from the underlying Sigma protocols, which we dub *instance-dependent malleability*, where it should be possible to maul the prover’s first message in such a way that we can later balance the prover’s last message as well as the verifier’s challenge.

For the RF in Fig. 6, we prove the following result, whose proof appears in the full version [29] of this paper.
Theorem 6. Let \( \Sigma_0 = (P^0 = (P^0_1, P^0_2), V^0) \) and \( \Sigma_1 = (P^1 = (P^1_1, P^1_2), V^1) \) be instance-dependent malleable Sigma protocols with unique responses, for relations \( R_0 \) and \( R_1 \). The RFW of Fig. 6 preserves completeness, and is weakly HVZK/WI preserving for the prover of the Sigma protocol \( \Sigma_{\text{OR}} \) for relation \( R_{\text{OR}} \).
5 Previous Work

5.1 Comparison with Mironov and Stephens-Davidowitz

In their original paper, Mironov and Stephens-Davidowitz [40] build RFs for arbitrary two-party protocols. While their results are related to ours, since IPSes are just a special case of two-party computation, there are some crucial differences which we highlight below.

The first RF construction sanitizes a specific combination of re-randomizable garbled circuits and oblivious transfer, for obtaining general-purpose private function evaluation. The second RF construction sanitizes any two-party protocol, at the price of encrypting the full transcript under public keys that are broadcast at the beginning of the protocol. Both constructions can be instantiated based on (variants of) the DDH assumption. When cast to IPSes, their results yield:

(i) A RF for the prover that weakly preserves ZK. This is comparable to our RF achieving weak ZK preservation using malleable Sigma protocols and key-malleable commitments. However, our constructions have the advantage that we do not need to change the initial IPS, and thus our RF can be applied directly to already existing implementations in a fully transparent manner (and without introducing any overhead).

(ii) A RF for the prover satisfying a property called strong exfiltration resistance against an eavesdropper, which means that exfiltration resistance holds w.r.t. an arbitrarily subverted prover talking to the honest verifier. Note that the latter does not contradict our impossibility result ruling out strong ZK preservation, as our attacks crucially rely on the fact that the distinguisher can (passively) corrupt the verifier.

(iii) A RF for the verifier satisfying both strong exfiltration resistance and the following weak guarantee: No malicious prover can find statements $x_0, x_1$ such that it can distinguish transcripts obtained by talking to an arbitrarily subverted verifier holding either input $x_0$ or input $x_1$. Note that the latter does not contradict our impossibility result that rules out weak soundness preservation, since none of the above guarantees imply soundness preservation.

We observe that the above results have at least one of the following drawbacks: (i) The RF is not transparent, i.e. it cannot be applied to the initial protocol as is; (ii) The resulting sanitized protocol is not efficient, as we first need to encode the function being computed as a circuit.

Our techniques allow to overcome these limitations in the concrete case of IPSes, as our RFs are both transparent (i.e. they can be applied directly to already deployed protocols) and efficient (i.e. the sanitized IPSes have exactly the same efficiency as the original, both in terms of round and communication complexity). We see this as the main novelty of our work.

5.2 Additional Related Works

Besides the already mentioned constructions, RFs have also been realized in other settings including digital signatures [5], secure message transmission and key exchange [21, 12], and oblivious transfer [40, 12].

Moreover, a few other lines of research recently\(^\text{10}\) emerged to tackle the challenge of protecting cryptographic algorithms against (different forms of) subversion. We review the

\(^{10}\text{All these research directions have their roots in the seminal works of Young and Yung [52] and Simmons [50], in the settings of kleptography and subliminal channels.}\)
main ones below.

Algorithm substitution attacks
Bellare, Patterson, and Rogaway [11] studied subversion of symmetric encryption schemes in the form of algorithm substitution attacks (ASAs). In particular, they show that undetectable subversion of the encryption algorithm is possible, and may lead to severe security breaches; moreover, they prove that deterministic, stateful, ciphers are secure against the same type of ASAs. Follow-up works improved the original paper in several aspects [18, 10], and explored the power of ASAs in other contexts, e.g. digital signatures [5], secret sharing [31], and message authentication codes [3].

Backdoors
Another form of subversion consists of all those attacks that surreptitiously generate public parameters (primes, curves, etc.) together with secret backdoors that allow to bypass security. The study of this type of subversion is motivated by the DUAL_EC_DRBG PRG incident.
A formal study of parameters subversion has been considered for several primitives, including pseudorandom generators [20, 19], hash functions [27], non-interactive zero knowledge [8], and public-key encryption [6].

Cliptography
Russell et al. [46] (see also [47, 4]) consider a different approach to the immunization of cryptosystems against complete subversion (i.e., when all algorithms can be subverted by the attacker): offline/online black-box testing. This amounts to introducing an external entity, called the watchdog, whose goal is to test, either in an online or in an offline fashion, whether a given cryptographic implementation is compliant with its specification.
Hence, a cryptosystem is deemed secure against complete subversion if there exists a universal watchdog such that, for every attacker subverting all algorithms, either the watchdog detects subversion with high probability, or the cryptoscheme remains secure even when using its subverted implementation.

Self-guarding
Yet another approach towards thwarting subversion is that of self-guarding [28]. The idea here is to assume a trusted initialization phase in which the honest parties possess a genuine implementation of the cryptosystem, before subversion takes place. This phase is used in order to generate samples that will be exploited later, together with additional simple operations that need to be implemented from scratch, to prevent leakage in the face of subversion attacks.

6 Conclusion
We showed how to design cryptographic reverse firewalls allowing to preserve security of interactive proof systems in the face of subversion. Our firewalls apply to a large class of Sigma protocols meeting a natural malleability property, and can be extended to cover classical applications of Sigma protocols for designing zero-knowledge proofs and for proving compound statements.
We leave it as an intriguing open problem to design a reverse firewall for the OR composition of Sigma protocols that are delayed input, as considered in [13, 14].
References


Cryptographic Reverse Firewalls for Interactive Proof Systems


Alessandra Scafuro and Ivan Visconti. On round-optimal zero knowledge in the bare public-key model. In David Pointcheval and Thomas Johansson, editors, EUROCRYPT 2012, volume 7237

