On the Incoercibility of Digital Signatures

Ashley Fraser
Department of Computer Science
University of Surrey
Guildford, UK
a.fraser@surrey.ac.uk

Lydia Garms
Keyless Technologies Limited
London, UK
lydia.garms@keyless.io

Elizabeth A. Quaglia
Information Security Group
Royal Holloway, University of London
Egham, UK
elizabeth.quaglia@rhul.ac.uk

2023 IEEE 36th Computer Security Foundations Symposium (CSF) | 979-8-3503-2192-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSF57540.2023.00018

***Abstract*—We introduce incoercible digital signature schemes, a variant of a standard digital signature. Incoercible signatures enable signers, when coerced to produce a signature for a message chosen by an attacker, to generate fake signatures that are indistinguishable from real signatures, even if the signer is compelled to reveal their full history (including their secret signing keys and any randomness used to produce keys/signatures) to the attacker. Additionally, we introduce an authenticator that can detect fake signatures, which ensures that coercion is identified. We present a formal security model for incoercible signature schemes that comprises an established definition of unforgeability and captures new notions of weak receipt-freeness, strong receipt-freeness and coercion-resistance. We demonstrate that an incoercible signature scheme can be viewed as a transformation of any generic signature scheme. Indeed, we present two incoercible signature scheme constructions that are built from a standard signature scheme and a sender-deniable encryption scheme. We prove that our first construction satisfies coercion-resistance, and our second satisfies strong receipt-freeness. We conclude by presenting an extension to our security model: we show that our security model can be extended to the designated verifier signature scheme setting in an intuitive way as the designated verifier can assume the role of the authenticator and detect coercion during the verification process.**

***Index Terms*—digital signatures, incoercibility, receipt-freeness, coercion-resistance**

I. Introduction

The fundamental security property of a digital signature scheme is considered to be the property of existential unforgeability against an adaptive chosen-message attack [1]. This captures the property that an attacker, with access only to public information, cannot output a valid message/signature pair (i.e., a forgery). In recognition of the threat posed by attackers with stronger, potentially coercive, capabilities, several security notions have been proposed that strengthen the traditional security model for signature schemes (cf. §I-A). These notions have predominantly focused on *key exposure attacks* [2]–[9], whereby a signer is coerced to reveal (part of) their secret key to an attacker. The proposed schemes generally allow the signer to recover from the attack, most commonly by updating their secret key. The security of such schemes requires that the attacker cannot produce a valid signature on behalf of a signer whose key has been exposed. Other works have focused on specific techniques that allow a signer to evade coercion [10], [11]. These works consider an attacker that requests a signer to produce a signature for a particular message. The proposed solutions introduce a trusted authority that can detect coercion, and their security model requires that the attacker cannot distinguish a signer evading coercion from a signer who cooperates with the attacker.

In this work, we introduce *incoercible signatures*, which follow the approach of [10], [11] in that coercion can be detected by a trusted authority, the authenticator. Our contributions distinguish themselves from the existing literature in the following way. Firstly, we preserve the essence of a standard signature scheme, i.e., our primitive is fully non–interactive, during key generation, signing and verification. Secondly, while previous works modelled signers revealing their secret key, we allow the attacker to demand the *full* transcript of the signer throughout the protocol, thereby including, along with the secret key, any randomness chosen in key generation and signing. Thirdly, we recognise the benefits of capturing several variants of incoercibility and, therefore, our security model also captures attackers that demand a full transcript only *after* having provided instructions to the signers.

A. Existing Work on Incoercibility

The concept of incoercibility first emerged in the context of electronic voting (e-voting) where it was formalised as a hierarchy [12] of two security properties: receipt-freeness [13] and coercion-resistance [14]. Additionally, simulation-based definitions of incoercibility have been put forth in the generic multi-party protocol setting [15]–[17], again defined as a hierarchy of the receipt-freeness and coercion-resistance properties.

In the digital signatures literature, coercion has been addressed in several ways. Security models and schemes have been proposed to protect against (partial) key exposure attacks, and mechanisms have been introduced to enable the signer to warn an authority of such attacks, either via direct communication or by embedding a coercion warning into the signature. We discuss these approaches in more detail next.

The work of Ashley Fraser was supported by the EPSRC and the UK government as part of the Centre for Doctoral Training in Cyber Security at Royal Holloway, University of London under grant number EP/P009301/1, and the EPSRC Next Stage Digital Economy Centre in the Decentralised Digital Economy (DECaDE) at the University of Surrey under grant number EP/T022485/1.
The work of Lydia Garms was supported by the Innovate UK funded project AQuaSec, whilst working at Royal Holloway, University of London, and a research grant from Nomadic Labs and the Tezos Foundation, whilst working at the IMDEA Software Institute.

DOI 10.1109/CSF57540.2023.00018

a) Key Exposure Attacks: An attacker that can obtain the secret key of a signer can easily construct valid forgeries on behalf of the signer. This is known as a key exposure attack and several solutions that allow recovery from such an attack exist. The earliest solutions required key distribution [18]–[20], but the most prevalent solution allows the signer to update their secret key [2]–[4], [6]–[8]. This approach became prominent following the introduction of forward secure signatures [2], which guarantee that an attacker cannot produce valid forgeries for any time period prior to key update. Subsequent works, for example [3], [4], [6]–[8], extend this guarantee, providing an assurance that an attacker cannot produce valid forgeries for time periods prior to, and following, the key update. However, many of these solutions provide these guarantees for partial key exposure only [3], [4], [7], [8]. By contrast, [2], [6], [8] consider full key exposure attacks. In a departure from this approach, monotonic signature schemes [9] allow the signer, in response to a partial key exposure attack, to update the verification algorithm, rather than the secret signing key. Monotonic signatures ensure that the attacker cannot produce valid forgeries after the verification algorithm is updated, but forgeries created before the algorithm update are valid. In this work, we consider an attacker that demands a full transcript from the signer, and require that the attacker cannot distinguish cooperating and non-cooperating signers. Additionally, our strongest security definition guarantees that an attacker cannot produce valid forgeries on behalf of a coerced signer.

b) Communicating Key Exposure: Generally, signature schemes that are secure against key exposure attacks present a limited window of time within which the attacker can produce a valid forgery, e.g., the time period when key exposure occurs in the case of [3], [4], [6]–[8]. Moreover, the requirement to update keys/algorithms may alert an attacker to an unsuccessful key exposure attack. For example, in the case of monotonic signatures, the attacker can check if their forgery is valid. If the forgery is invalid, it is clear that the signer updated their verification algorithm and thus evaded coercion. Similar arguments hold for other schemes secure against key exposure attacks. To address this, key evolving signature schemes [8] introduced a trusted authority with whom the signer can communicate. The signer is required to update their keys at regular intervals; if the signer does not contact the authority before key update occurs, any signatures generated since the previous key update period become invalid. In this way, forgeries created by the attacker can be invalidated. Moreover, the signer can communicate to the authority that the signatures were coerced. In doing so, the signer can evade coercion and, if the authority does not publicly invalidate the signatures, the attacker is not alerted to the unsuccessful coercion. Similarly, funkspiel schemes [5] introduce a trusted authority with whom the signer can communicate the compromise of keys. In our work, we also introduce a trusted authority that can detect coercion, though we do not require that the signer communicates with the authority. Rather, we include a 'coercion alert' in the signature that can be recovered only by the trusted authority.

c) Embedded Secret Signature Schemes: Embedded secret signatures [10], [11] allow signers to produce signatures that contain an embedded warning that can be extracted from the signature only by a trusted authority. In this way, the signer can evade coercion to sign a message chosen by the attacker, without detection by the attacker. Incoercible signatures achieve a similar goal. A key difference, however, is that embedded secret signature schemes require that the trusted authority interact with every signer during key generation. By contrast, incoercible signature schemes are non-interactive. Moreover, the security model for an embedded secret signature requires that an attacker cannot determine whether the signer signed the attacker's message or not, when provided with the signer's secret key. In comparison, our security model for incoercibility considers an attacker with access to full signer transcripts that include randomness used during signing and key generation, in addition to the signer's secret key.

d) Deniability and Incoercibility: Deniability is a property that can be useful in the context of coercion within cryptographic protocols. Indeed, if a protocol participant is coerced to perform a particular action, deniability can be used to allow the participant to perform a different action (or no action) and prove that they followed the coercer's instructions. However, deniability captures a wider range of attacks. The participant can, when coerced not to do something, perform the action and deny it afterwards. In other words, deniability captures the ability of a participant to repudiate their actions to all but the intended receiver.

Deniability has been considered widely in the cryptographic literature (e.g., [21]–[25]). In the context of digital signatures, deniability, if defined analogously to other cryptographic primitives, would provide an assurance that the signer can deny having produced a signature. Therefore, deniability captures coercive attacks where the attacker instructs the signer *not* to sign messages. However, deniability in this context seems difficult to achieve. Certainly, a message/signature pair is public, which suggests that a signer cannot deny producing the signature. Therefore, in this work, we focus on incoercibility, which captures coercive attacks where the attacker instructs the signer to sign a message of their choice. Additionally, we introduce incoercible strong designated verifier signature schemes, which are not publicly verifiable, capturing incoercibility as well as the aspects of deniability featured in strong designated verifier schemes.

B. Our Contributions

The main contribution of this paper is the introduction and formalisation of incoercible signature schemes. In particular, we define three notions of incoercibility: weak receipt-freeness, strong receipt-freeness and coercion-resistance. We demonstrate the feasibility of our primitive by presenting constructions of coercion-resistant and strong receipt-free incoercible signature schemes, and we prove that they satisfy our security model. We also consider a relevant related primitive: designated verifier signature schemes. Finally, we motivate the study of incoercible

signatures by describing some applications. We next describe our contributions in detail.

1) Defining Incoercible Signatures: We introduce the syntax for an incoercible signature scheme (§II). Incoercibility means that a signer can sign messages as they desire and, conversely, need not produce signatures for messages that they do not want to sign. We define incoercible signatures as a variant of a standard signature scheme, providing algorithms for key generation, signing and public verification of a signature. To realise incoercible signatures, we introduce two new requirements into our syntax. Firstly, we introduce the ability to produce 'fake' signatures that are indistinguishable from real signatures. In this way, a coerced signer can produce a fake signature for a message chosen by the attacker. Formally, we introduce a new algorithm FakeSign that outputs a fake signature. Secondly, we introduce a trusted authority, which we call the authenticator, that can distinguish real signatures produced by running standard signing algorithm Sign and signatures that are the output of algorithm FakeSign. In this paper, we adopt the convention of calling a signature valid or verifiable if it passes public verification, and, additionally, call a signature authentic if the designated authenticator deems the signature to be real, as opposed to fake. In our syntax, we introduce an algorithm Authenticate that is run by the authenticator and outputs whether the signature is authentic or not. Note that, without a trusted authority, fake signatures are identical to real signatures. However, an authenticator is only trusted to determine whether a signature is coerced and learns nothing about the signer's secrets. Crucially, this means that the authenticator cannot produce forgeries on behalf of any signer. We also highlight that, in practice, the authenticator must be carefully chosen to ensure that the correct authority is alerted to coercion.

Our syntax assumes that signers keep a transcript of their actions, which we denote trans. Informally, trans is a record of all computations performed, and inputs generated, by the signer. We include such a transcript in our syntax to model the fact that a signer may present an attacker with 'evidence' (i.e., the transcript) to prove that they have followed the attacker's instructions. We stress that our syntax includes the transcript only for the purpose of modelling coercive attacks. It is not necessary for a signer to keep track of the transcript in order to sign. In this way, incoercible signatures can be distinguished from stateful signatures, which require that the signer's state is input to the signing algorithm to compute a signature. A transcript is uniquely defined by a given signature scheme, and we provide a concrete example of this here. We illustrate the form of trans in the case of the DSA/ECDSA signature scheme, whose description is taken from [26].

Example 1 (DSA/ECDSA)**.** *DSA (respectively, ECDSA) is defined relative to a set of public parameters* $pp_{\mathsf{SIG}} = (\mathbb{G}, g, q)$ *where* $\mathbb{G}$ *is a* q*-order subgroup of* $\mathbb{Z}_p^*$ *(resp.,* $E(\mathbb{Z}_p)$*) for* p *a prime. We denote the generator of* $\mathbb{G}$ *as* g*. We also assume that functions* $G : \mathbb{G} \to \mathbb{Z}_q$ *and* $H : \{0,1\}^* \to \mathbb{Z}_q$ *are defined during setup. Then, DSA/ECDSA is defined in Figure 1. The transcript* trans *(after key generation and the signing of one message* m*) is defined as the tuple* $((pk_{\mathsf{SIG}}, sk_{\mathsf{SIG}}), (k, \sigma, m))$*. We note that, with the transcript, it is possible to verify that* $pk_{\mathsf{SIG}} = g^{sk_{\mathsf{SIG}}}$ *and recompute the signature* σ *generated by the signer.*

$\mathsf{KGen}(pp_{\mathsf{SIG}})$

$sk_{\mathsf{SIG}} \leftarrow \mathbb{Z}_q;\ pk_{\mathsf{SIG}} \coloneqq g^{sk_{\mathsf{SIG}}};$

return $(pk_{\mathsf{SIG}}, sk_{\mathsf{SIG}})$

$\mathsf{Sign}(pp_{\mathsf{SIG}}, sk_{\mathsf{SIG}}, m)$

$k \leftarrow \mathbb{Z}_q^*;\ r \coloneqq G(g^k);$

$s \coloneqq (k^{-1} \cdot (H(m) + sk_{\mathsf{SIG}} \cdot r)) \mod q;$

return $\sigma = (r, s)$

$\mathsf{Vf}(pp_{\mathsf{SIG}}, pk_{\mathsf{SIG}}, m, \sigma)$

if $r \neq G(g^{H(m)\cdot s^{-1}} \cdot pk_{\mathsf{SIG}}^{r\cdot s^{-1}})$ **return** 0

return 1

Fig. 1: The DSA/ECDSA signature scheme for which we define the transcript in Example 1.

Additionally, our syntax models the generation of a fake transcript, denoted $\mathsf{trans}_{\mathsf{fake}}$. The fake transcript is used by a signer that attempts to evade coercion by providing a transcript that will convince an attacker that the signer followed the attacker's instructions. As we shall see in Section II-A, maintaining a real and fake transcript is necessary to ensure that a signer can evade coercion, the main security goal of incoercible signatures.

2) A Security Model for Incoercible Signatures: We present a comprehensive security model for incoercible signatures (§II-A). Our security model captures standard notions of correctness and unforgeability for a signature scheme, adapted to our syntax. We also introduce a further property that we call completeness, which captures the notion that an honestly generated signature is authentic. We then define three further security properties for incoercible signatures: weak receipt-freeness, strong receipt-freeness and coercion-resistance. In this way, our security model captures three variants of incoercibility that are influenced and inspired by the existing literature [16], [17], [27], [28]. Here we provide an overview of our incoercibility notions, and present our formal definitions in Section II-A.

Our incoercibility notions consider signers that fall into one of the following three categories: 1) *Honest* signers follow the protocol and do not interact with the attacker; 2) *Corrupt* signers are controlled by the attacker. The attacker can act on behalf of corrupt signers and is assumed to have the secret signing key of corrupt signers; 3) *Coerced* signers are signers that are provided with instructions by the attacker. The attacker can request that the signer provide some 'proof' that they followed the attacker's instructions. Coerced signers may appear to cooperate with the attacker when they are, in fact, deceiving

the attacker. For incoercible signatures, we are most interested in this category of signers.

Our security definitions allow an attacker to *adaptively* corrupt or coerce signers, meaning that corruption or coercion can occur at any point after key generation. Furthermore, we assume that signers can fall into only one of these categories. That is, corrupt signers cannot be coerced and coerced signers cannot be corrupt. We omit simultaneous corruption and coercion from our security model for the following reasons. Firstly, in the former scenario, an attacker gains no advantage by coercing a corrupt signer. Secondly, in the latter scenario, coercion can be trivially detected if the full transcript of the signer is revealed upon corruption. Indeed, in [17], it is established that a notion of incoercibility that allows simultaneous corruption and coercion is, arguably, too strong a notion and, for most application scenarios, is unnecessary.

Our notions of incoercibility are grounded in the understanding of incoercibility developed over the years, in particular in the e-voting literature [14], [27], [28]. We reflect this understanding by introducing the properties of *indistinguishability* and *soundness* to our security model, as we describe next.

a) Indistinguishability: We introduce three variants of indistinguishability that capture the following intuition: an attacker, who can request transcripts from coerced signers, cannot distinguish a signer that cooperates with an attacker and presents a real transcript from a signer that evades coercion and presents a fake transcript. Our weakest notion of indistinguishability, which we call IND1, considers an attacker that interacts with signers only after signing. Building on this, we define IND2 indistinguishability, which captures an attacker that can interact with signers throughout the protocol (i.e., before and after signing). Finally, our strongest variant, IND3 indistinguishability, captures an attacker that provides the signer with the randomness used to sign, in addition to the attacker's message.

b) Soundness: It is essential that, if a signature is fake, the authenticator will determine the signature to be inauthentic. To capture this, we introduce *soundness*, which guarantees that a coerced signer can communicate coercion to the authenticator. It requires that, if a signature is the output of algorithm FakeSign, the authenticator will determine the signature to be fake. We also recognise that some attackers may attempt to act on behalf of signers (for example, by attempting to produce signatures on behalf of coerced signers). Consequently, we introduce a second soundness property that we call *strong soundness*. Strong soundness is the property that an attacker cannot output an authentic signature on behalf of a coerced, but uncooperative, signer. In other words, strong soundness requires that an adversary, who can request fake signatures and fake transcripts on behalf of coerced signers, cannot produce an authentic signature for a coerced signer.

Our indistinguishability and soundness properties can be combined in various ways to capture a spectrum of incoercibility attackers. In this work, we focus on combining these properties to achieve three well-established incoercibility notions from the literature. Following convention in the e-voting literature, we call our incoercibility properties weak receipt-freeness, strong receipt-freeness and coercion-resistance, described in more detail in Section II.

3) Achieving Incoercible Signatures: Suprisingly, we show that even the strongest notion of incoercibility for digital signatures (cf. §II-B) can be achieved using existing well-established cryptographic primitives. Indeed, we present two conceptually simple incoercible signature scheme constructions (§III) that rely on a generic standard signature scheme and a sender-deniable encryption scheme. We prove that our first construction satisfies coercion-resistance and our second construction satisfies strong receipt-freeness, providing proof sketches in Section III and full proofs in the full version of this paper [29]. By presenting such constructions, we demonstrate that our notion of coercion-resistance is achievable, and that our notion of strong receipt-freeness is achievable with some efficiency savings (when compared to our coercion-resistant construction). Note that we do not present a construction for our weak receipt-freeness notion. Of course, our two constructions satisfy weak receipt-freeness, but we do not present a construction that satisfies *only* weak receipt-freeness. We do this as we aim to achieve the strongest security possible in our constructions (whilst balancing this with efficiency). We leave as an open problem whether more efficient constructions are possible that can satisfy our weakest notion. Looking forward, in Section IV, we explicit the relation between incoercibility and deniable encryption, and demonstrate that weak receipt-freeness implies a variant of deniable encryption. As a result, we conjecture that the efficiency of weak receipt-free schemes and deniable encryption are closely related. That being said, we choose to include weak receipt-freeness in our security model to provide a complete vision of incoercibility for the digital signature space.

4) Extending Incoercibility to the Designated Verifier Setting: We demonstrate that our security notions can be extended in an intuitive way to the setting of designated verifier signature schemes (§V). Indeed, designated verifier signature schemes are well-suited to the notion of incoercibility because the designated verifier can assume the role of the authenticator and detect coercion during verification. We provide a security model for incoercible strong designated verifier signature schemes that captures notions of correctness, unforgeability, source hiding and privacy of signer's identity as defined in [30]. Additionally, we define incoercibility, which corresponds to the definitions of weak receipt-freeness, strong receipt-freeness and coercion-resistance for publicly verifiable signature schemes. We also present a construction that satisfies our security model for coercion-resistance. Our construction is similar to our coercion-resistant construction, except that the standard publicly verifiable signature scheme is replaced with a strong designated verifier signature scheme. We present an overview of our strong designated verifier signature results in Section V and provide full details of our results in [29].

C. Utility of Incoercible Signatures

In a world where online intimidation is increasingly prevalent, it seems clear that extending the notion of incoercibility to the setting of digital signatures is a useful endeavour. Indeed, in the context of signatures with a designated verifier, it is straightforward to see that our notion of incoercibility can offer protection against forced influence or coercion: the coerced signer can signal the coercion attempt to the verifier, who is uniquely positioned to verify the signature and discard it.

With respect to standard digital signatures, coercion can only be detected by the authenticator and not when signatures are verified, making a successful evasion more challenging. Nevertheless, incoercible signatures can be of use. Consider, for instance, a scenario in which members of an organisation can produce endorsements on behalf of other members. Such endorsements can be used, for example, to recommend a particular member for an advertised role within the organisation. It may be desirable that members *publicly* endorse other members; endorsements accompanied with publicly verifiable signatures can facilitate this. However, it is also possible that a member may be coerced by another member (the coercer) to produce an endorsement. To address this, the organisation can implement a coercion-resistant incoercible signature scheme, ensuring that the coerced member can indicate coercion to an authenticator (i.e., the entity who ultimately determines which members are appointed to advertised roles). The authenticator can subsequently refuse to appoint the coercer to the role, without revealing the fact that the endorsing member evaded the coercion attempt. As such, the use of coercion-resistant incoercible signatures achieves a balance of public verifiability and protection from coercive attacks.

Similarly, in reputation systems reviews are provided for a product and can be accompanied by a publicly verifiable signature to provide trust in their origins, without relying on a trusted third party. Moreover, an authenticator can provide a tally that averages the uncoerced reviews, without revealing which reviews are included in this tally. In this scenario, a verifier has access to two sources of information: a set of scores that may be coerced but do not rely on a trusted third party and a tally that does not contain coerced reviews (assuming the authenticator behaves honestly). The verifier can choose which source to trust.

In these scenarios, the authenticator need not publicly reveal the coercion attempt. Indeed, in the first scenario, the authenticator could fabricate a reason for not appointing the coercer to the role. Along with the designated verifier setting, this is crucial to the utility of incoercible signatures. As with other incoercible primitives, e.g., deniable encryption, the user of an incoercible signature must consider the non-cryptographic consequences of evading coercion. In some scenarios, the actions of the verifier may indicate an unsuccessful coercion attempt to the coercer. For example, if a coercer requests the transfer of funds, the sender can indicate coercion to the entity that will perform the transfer by producing an incoercible signature for the transfer request. The entity will be alerted to the coercion attack and will not transfer the funds. However, if the coercer does not receive the funds, they will be alerted to the fact that the sender evaded the coercion attempt. In this scenario it is more difficult for the verifier to provide a convincing reason not to send the funds, and so incoercible signatures are not suitable.

Incoercible signatures can also be used within larger protocols for which receipt-freeness and coercion-resistance are desirable properties. One example of such a protocol is e-voting. Briefly, registered voters could sign an encrypted vote with an incoercible signature. Voter ballots (i.e., their encrypted vote and incoercible signature) can be submitted to a trusted election official for tallying. In the role of authenticator, the election official can check whether the signature is coerced and can discard any ballots that signal coercion. In this way, the election official can exclude ballots from the result without revealing any coercion attempts. Furthermore, ballots, including coerced ballots, can be posted to a public bulletin board. Using an incoercible signature means that anyone can verify signatures attached to ballots, thus checking that all ballots are submitted by registered voters.

II. Incoercible Signatures

In this section, we introduce the syntax and security model for an incoercible signature scheme. Recall from the introduction that incoercible signatures are a variant of a standard signature scheme, providing standard algorithms for key generation, signing and public verification of a signature. Additionally, incoercible signatures are equipped with an algorithm $\mathsf{FakeSign}$ that allows signers to generate fake signatures and algorithm $\mathsf{Authenticate}$ that is run by the authenticator and outputs whether the signature is authentic (i.e. a real signature that is the output of algorithm Sign) or not.

Our syntax assumes that signers maintain a transcript of their actions, which we denote trans. We define the transcript to contain all secrets and randomness generated by the signer during key generation and signing (cf. Example 1). We also introduce a fake transcript $\mathsf{trans}_{\mathsf{fake}}$ into our syntax. Formally, we define the syntax for an incoercible signature scheme in Definition 1.

Definition 1 (Incoercible Signature Scheme)**.** *Let* trans *denote the transcript of a signer that contains all secrets and randomness generated by the signer during key generation and signing. Then, an* incoercible signature scheme $\mathsf{INC\text{-}SIG}$ *is a tuple of probabilistic polynomial time (PPT) algorithms* $(\mathsf{Setup}, \mathsf{AKGen}, \mathsf{SKGen}, \mathsf{FakeTrans}, \mathsf{Sign}, \mathsf{FakeSign}, \mathsf{Verify}, \mathsf{Authenticate})$ *such that:*

- $\mathsf{Setup}(1^\lambda)$ On input of security parameter 1^λ, Setup outputs public parameters pp.
- $\mathsf{AKGen}(pp)$ On input of public parameters pp, AKGen outputs an authenticator key pair (pk_A, sk_A) where pk_A is the authenticator's public key and sk_A is the authenticator's private key.
- $\mathsf{SKGen}(pp, pk_A)$ On input of public parameters pp and authenticator public key pk_A, SKGen outputs a signer

key pair (pk_S, sk_S) where pk_S is the signer's public key and sk_S is the signer's private key, and an initial transcript trans.

$\mathsf{FakeTrans}(pp, pk_A, \mathsf{trans}, \mathsf{trans}_{\mathsf{fake}})$ On input of public parameters pp, authenticator public key pk_A, a transcript trans and a fake transcript $\mathsf{trans}_{\mathsf{fake}}$, FakeTrans outputs an updated fake transcript $\mathsf{trans}_{\mathsf{fake}}$. We write that $\mathsf{trans}_{\mathsf{fake}} \leftarrow \mathsf{FakeTrans}(pp, pk_A, \mathsf{trans}, \perp)$ generates an initial fake transcript.

$\mathsf{Sign}(pp, sk_S, pk_A, m, \mathsf{trans})$ On input of public parameters pp, signer secret key sk_S, authenticator public key pk_A, message m and transcript trans, Sign outputs a signature σ and an updated transcript trans.

$\mathsf{FakeSign}(pp, sk_S, pk_A, m, \mathsf{trans}, \mathsf{trans}_{\mathsf{fake}})$ On input of public parameters pp, signer secret key sk_S, authenticator public key pk_A, message m, a transcript trans and a fake transcript $\mathsf{trans}_{\mathsf{fake}}$, FakeSign outputs a signature σ and an updated fake transcript $\mathsf{trans}_{\mathsf{fake}}$.

$\mathsf{Verify}(pp, pk_S, m, \sigma)$ On input of public parameters pp, signer public key pk_S, message m and signature σ, Verify outputs 1 if σ verifies, and 0 otherwise.

$\mathsf{Authenticate}(pp, pk_S, sk_A, m, \sigma)$ On input of public parameters pp, signer public key pk_S, authenticator secret key sk_A, message m and signature σ, Authenticate outputs 1 if σ is authentic, and 0 otherwise.

We require that an incoercible signature scheme satisfies *correctness*, which, as for standard signature schemes, requires that honestly generated signatures verify. Additionally, an incoercible signature scheme must satisfy *completeness*, the property that honestly generated signatures are authentic.

Definition 2 (Correctness)**.** *An incoercible signature scheme* INC-SIG *satisfies* correctness *if, for any message* $m \in \{0,1\}^*$*, there exists a negligible function* negl *such that*

$$\Pr\left[\begin{array}{l} pp \leftarrow \mathsf{Setup}(1^\lambda); \\ (pk_A, sk_A) \leftarrow \mathsf{AKGen}(pp); \\ (pk_S, sk_S, \mathsf{trans}) \leftarrow \mathsf{SKGen}(pp, pk_A); \\ (\sigma, \mathsf{trans}) \leftarrow \mathsf{Sign}(pp, sk_S, pk_A, m, \mathsf{trans}) \end{array} : \mathsf{Verify}(pp, pk_S, m, \sigma) = 1\right] \geq 1 - \mathsf{negl}(\lambda).$$

Definition 3 (Completeness)**.** *An incoercible signature scheme* INC-SIG *satisfies* completeness *if, for any message* $m \in \{0,1\}^*$*, there exists a negligible function* negl *such that*

$$\Pr\left[\begin{array}{l} pp \leftarrow \mathsf{Setup}(1^\lambda); \\ (pk_A, sk_A) \leftarrow \mathsf{AKGen}(pp); \\ (pk_S, sk_S, \mathsf{trans}) \leftarrow \mathsf{SKGen}(pp, pk_A); \\ (\sigma, \mathsf{trans}) \leftarrow \mathsf{Sign}(pp, sk_S, pk_A, m, \mathsf{trans}) \end{array} : \begin{array}{l} \mathsf{Authenticate}(pp, pk_S, sk_A, \\ m, \sigma) = 1 \end{array}\right] \geq 1 - \mathsf{negl}(\lambda).$$

A. Security Model

We present a security model for our syntax capturing a standard definition of existential unforgeability against a chosen message attack (EUF-CMA) [1], adapted to the setting of incoercible signature schemes. In our EUF-CMA experiment, we require that an adversary cannot output a valid signature on behalf of a signer, where the signature is not the output of the signing oracle, even if the adversary has access to the authenticator's key pair.

Definition 4 (Unforgeability)**.** *An incoercible signature scheme* INC-SIG *satisfies* existential unforgeability against a chosen message attack (EUF-CMA) *if there exists a negligible function* negl *such that*

$$\Pr\left[\begin{array}{l} \mathsf{Query} \leftarrow \emptyset; \\ pp \leftarrow \mathsf{Setup}(1^\lambda); \\ (pk_A, sk_A) \leftarrow \mathsf{AKGen}(pp); \\ (pk_S, sk_S, \mathsf{trans}) \leftarrow \mathsf{SKGen}(pp, pk_A); \\ (m^*, \sigma^*) \leftarrow \mathcal{A}^{\mathsf{SIGN}(\cdot)}(pp, pk_S, pk_A, sk_A) \end{array} : \begin{array}{l} \mathsf{Verify}(pp, pk_S, m^*, \sigma^*) \\ = 1 \\ \wedge\ m^* \notin \mathsf{Query} \end{array} \leq \mathsf{negl}(\lambda)\right]$$

where oracle $\mathsf{SIGN}(m)$*, on input message* m*, updates set* Query *to include message* m*, computes* $(\sigma, \mathsf{trans}) \leftarrow \mathsf{Sign}(pp, sk_S, pk_A, m, \mathsf{trans})$ *and outputs* σ*.*

We now present definitions of security properties that are intended to capture the fact that a signer, when coerced to sign a message by an attacker, can evade coercion, even if the signer is compelled to reveal their entire transcript to the attacker. In particular, we define two variants of receipt-freeness: weak and strong receipt-freeness. We complete our security model with a definition of coercion-resistance for incoercible signature schemes. Our definitions capture adaptive corruption and coercion strategies. We define the oracles for our experiments in Figure 2 and our experiments in Figure 3.

Our formal security definitions rely on a number of oracles. We write $\mathsf{X}_{(y_1,\ldots,y_n)}(z_1,\ldots,z_n)$ to denote oracle X that has access to parameters and lists $y_1,\ldots,y_n$ and takes as input $z_1,\ldots,z_n$. Oracle ADDU models key generation and the creation of an initial fake transcript for signers, outputting the signer's public key to the adversary. Oracle SIGN outputs an honestly generated signature and updates the fake transcript of the signer. We define oracle SIGN to take as input *any* signer (recall from the introduction that signers can be honest, coerced or corrupt) that has been previously input to oracle ADDU. Accordingly, our security model captures the fact that coerced signers, in addition to following instructions provided by the attacker, may produce honestly generated signatures for messages of their choosing. The adversary may call oracle CRPT to adaptively corrupt a signer, on the condition that the signer is not coerced, and obtain the secret key and transcript of the signer. Oracles CRCSIG, COERCE, FAKECOERCE and FAKESIGN adaptively coerce signers that are not previously corrupt and perform further functions on behalf of coerced signers. More specifically, oracle CRCSIG, depending on a bit b chosen in the security experiment, returns a real or fake signature for a message provided by the adversary. Similarly, oracle STCRCSIG returns a real or fake signature, and, allows the adversary to provide the randomness used to sign the message as input to the oracle. Oracle COERCE returns a real or fake transcript, depending on the bit b. Oracle FAKECOERCE reveals the fake transcript of a coerced signer. Finally, coerced signers can be input to FAKESIGN to obtain fake signatures. In our oracles and the corresponding experiments, we write $\mathbf{pk_S}$ to be the vector of signer public keys and $\mathbf{pk_S}[id]$ as

the public key pk_S of signer id. We define the secret keys, transcripts and fake transcripts of signers analogously.

1) Weak Receipt-Freeness: Broadly, weak receipt-freeness captures an attacker that coerces a signer to sign messages and, afterwards, demands evidence of the signer's cooperation. In our syntax, evidence refers to the signer's transcript. We do not consider that a weak receipt-freeness attacker may attempt to generate signatures on behalf of a coerced signer. Therefore, we require only a basic soundness requirement. Our definition of weak receipt-freeness captures two properties: IND1 indistinguishability and soundness.

IND1 *indistinguishability* defines an adversary that can request signatures from a coerced signer via oracle CRCSIG. Depending on a bit b chosen in the experiment, CRCSIG models a coerced signer that runs algorithm Sign or FakeSign to produce a signature. At the end of the experiment, the adversary can request transcripts of coerced signers via oracle COERCE, which returns a real or fake transcript, depending on the bit b. Additionally, the adversary can corrupt signers and request honestly generated signatures on behalf of any signer via oracles CRPT and SIGN respectively. We say that an incoercible signature scheme satisfies the indistinguishability requirement if the adversary can guess b with probability only negligibly more than $1/2$.

Soundness is a basic requirement requiring that, if a signature is the output of algorithm FakeSign, then algorithm Authenticate will output 1 with negligible probability.

Definition 5 (Weak Receipt-Freeness)**.** *An incoercible signature scheme* INC-SIG *satisfies* weak receipt-freeness *if the following conditions hold.*

- ***Indistinguishability*** (IND1)*: for any PPT adversary* $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$*, there exists a negligible function* negl *such that*

$$\Big|\Pr\Big[\mathsf{Exp}^{\mathsf{IND1},0}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)=1\Big] - \Pr\Big[\mathsf{Exp}^{\mathsf{IND1},1}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)=1\Big]\Big| \leq \mathsf{negl}(\lambda).$$

- ***Soundness****: for any message* $m \in \{0,1\}^*$*, there exists a negligible function* negl *such that*

$$\Pr\Big[\mathsf{Exp}^{\mathsf{sound}}_{\mathsf{INC\text{-}SIG}}(\lambda)=1\Big] \leq \mathsf{negl}(\lambda).$$

where $\mathsf{Exp}^{\mathsf{IND1},b}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)$ *for* $b \in \{0,1\}$ *and* $\mathsf{Exp}^{\mathsf{sound}}_{\mathsf{INC\text{-}SIG}}(\lambda)$ *are the experiments defined in Figure 3.*

2) Strong Receipt-Freeness: A strong receipt-freeness attacker can interact with signers throughout the protocol. Crucially, this means that a strong receipt-freeness attacker can demand the transcripts of coerced signers at any point. As such, our definition of strong receipt-freeness must capture IND2 indistinguishability. Additionally, strong receipt-freeness requires soundness, as defined for weak receipt-freeness, as we assume that attacker does not attempt to generate signatures on behalf of a coerced signer.

Our IND2 experiment is similar to IND1 with one key difference: in our IND2 indistinguishability experiment, the adversary can query oracle COERCE throughout the experiment. This models the fact that the attacker, rather than requesting transcripts of a coerced signer at the end of a protocol run, may demand the transcripts at any point. With respect to all other oracles queries, the IND2 indistinguishability adversary is identical to the IND1 indistinguishability adversary. We require that the adversary cannot determine whether the coerced signers are cooperating or evading coercion. Formally, as in our IND1 indistinguishability experiment, this means that the adversary can guess the bit b with probability only negligibly more than $1/2$.

Definition 6 (Strong Receipt-Freeness)**.** *An incoercible signature scheme* INC-SIG *satisfies* strong receipt-freeness *if the following conditions hold.*

- ***Indistinguishability*** (IND2)*: for any PPT adversary* $\mathcal{A}$*, there exists a negligible function* negl *such that*

$$\Big|\Pr\Big[\mathsf{Exp}^{\mathsf{IND2},0}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)=1\Big] - \Pr\Big[\mathsf{Exp}^{\mathsf{IND2},1}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)=1\Big]\Big| \leq \mathsf{negl}(\lambda).$$

- ***Soundness****: for any message* $m \in \{0,1\}^*$*, there exists a negligible function* negl *such that*

$$\Pr\Big[\mathsf{Exp}^{\mathsf{sound}}_{\mathsf{INC\text{-}SIG}}(\lambda)=1\Big] \leq \mathsf{negl}(\lambda).$$

where $\mathsf{Exp}^{\mathsf{IND2},b}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)$ *for* $b \in \{0,1\}$ *and* $\mathsf{Exp}^{\mathsf{sound}}_{\mathsf{INC\text{-}SIG}}(\lambda)$ *are the experiments defined in Figure 3.*

3) Coercion-Resistance: A coercion-resistance attacker controls coerced signers and can interact with signers throughout the protocol. Consequently, a coercion-resistant attacker can demand the transcripts of coerced signers at any point, and can attempt to produce signatures on their behalf. Our coercion-resistance definition captures IND3 indistinguishability and strong soundness.

Our IND3 experiment is identical to IND2 with the following exception. Rather than providing the adversary with access to oracle CRCSIG, the adversary can query oracle STCRCSIG, which allows the adversary to provide the randomness used to sign the message in addition to the message itself.

Strong soundness is defined with respect to an adversary that can request fake signatures and fake transcripts on behalf of coerced signers via oracles FAKESIGN and FAKECOERCE respectively. The adversary can also query oracle CRPT to corrupt signers, obtaining their transcripts and secret keys. Moreover, the adversary can request honestly generated signatures for any signer by calling oracle SIGN. For strong soundness, we require that the adversary cannot output an authentic signature on behalf of a coerced signer, where the signature is not the output of the signing oracle.

Definition 7 (Coercion-Resistance)**.** *An incoercible signature scheme* INC-SIG *satisfies* coercion-resistance *if the following conditions hold.*

$\underline{\mathsf{ADDU}_{(pp,pk_A,\mathbf{pk_S},\mathbf{sk_S},\mathsf{L},\mathsf{trans},\mathsf{trans_{fake}})}(id)}$
if $id \in \mathsf{L}$ **return** $\perp$
$\mathsf{L} \leftarrow \mathsf{L} \cup \{id\}$
$(\mathbf{pk_S}[id], \mathbf{sk_S}[id], \mathsf{trans}[id]) \leftarrow \mathsf{SKGen}(pp, pk_A)$
$\mathsf{trans_{fake}}[id] \leftarrow \mathsf{FakeTrans}(pp, pk_A, \mathsf{trans}[id], \perp)$
return $\mathbf{pk_S}[id]$

$\underline{\mathsf{COERCE}_{(\mathsf{L},\mathsf{corL},\mathsf{crcL},\mathsf{trans},\mathsf{trans_{fake}})}(id)}$
if $id \notin \mathsf{L} \setminus \mathsf{corL}$ **return** $\perp$
$\mathsf{crcL} \leftarrow \mathsf{crcL} \cup \{id\}$
if $b = 0$ **return** $\mathsf{trans}[id]$
if $b = 1$ **return** $\mathsf{trans_{fake}}[id]$

$\underline{\mathsf{CRPT}_{(\mathbf{sk_S},\mathsf{L},\mathsf{corL},\mathsf{crcL},\mathsf{trans})}(id)}$
if $id \notin \mathsf{L} \setminus \mathsf{crcL}$ **return** $\perp$
$\mathsf{corL} \leftarrow \mathsf{corL} \cup \{id\}$
return $\mathbf{sk_S}[id], \mathsf{trans}[id]$

$\underline{\mathsf{FAKECOERCE}_{(\mathsf{L},\mathsf{corL},\mathsf{crcL},\mathsf{trans_{fake}})}(id)}$
if $id \notin \mathsf{L} \setminus \mathsf{corL}$ **return** $\perp$
$\mathsf{crcL} \leftarrow \mathsf{crcL} \cup \{id\}$
return $\mathsf{trans_{fake}}[id]$

$\underline{\mathsf{SIGN}_{(pp,pk_A,\mathbf{sk_S},\mathsf{L},\mathsf{Query},\mathsf{trans},\mathsf{trans_{fake}})}(id, m)}$
if $id \notin \mathsf{L}$ **return** $\perp$
$(\sigma, \mathsf{trans}) \leftarrow \mathsf{Sign}(pp, \mathbf{sk_S}[id], pk_A, m, \mathsf{trans}[id])$
$\mathsf{trans_{fake}}[id] \leftarrow \mathsf{FakeTrans}(pp, pk_A, \mathsf{trans}[id], \mathsf{trans_{fake}}[id])$
$\mathsf{Query} \leftarrow \mathsf{Query} \cup \{(id, m)\}$
return σ

$\underline{\mathsf{CRCSIG}_{(pp,pk_A,\mathbf{sk_S},\mathsf{L},\mathsf{corL},\mathsf{crcL},\mathsf{trans},\mathsf{trans_{fake}})}(id, m)}$
if $id \notin \mathsf{L} \setminus \mathsf{corL}$ **return** $\perp$
$\mathsf{crcL} \leftarrow \mathsf{crcL} \cup \{id\}$
if $b = 0$ $(\sigma, \mathsf{trans}[id]) \leftarrow \mathsf{Sign}(pp, \mathbf{sk_S}[id], pk_A, m, \mathsf{trans}[id])$
if $b = 1$ $(\sigma, \mathsf{trans_{fake}}[id]) \leftarrow \mathsf{FakeSign}(pp, \mathbf{sk_S}[id], pk_A, m, \mathsf{trans}[id], \mathsf{trans_{fake}}[id])$
return σ

$\underline{\mathsf{FAKESIGN}_{(pp,pk_A,\mathbf{sk_S},\mathsf{L},\mathsf{corL},\mathsf{crcL},\mathsf{trans},\mathsf{trans_{fake}})}(id, m, r)}$
if $id \notin \mathsf{L} \setminus \mathsf{corL}$ **return** $\perp$
$\mathsf{crcL} \leftarrow \mathsf{crcL} \cup \{id\}$
$(\sigma, \mathsf{trans_{fake}}[id]) \leftarrow \mathsf{FakeSign}(pp, \mathbf{sk_S}[id], pk_A, m,$
$\mathsf{trans}[id], \mathsf{trans_{fake}}[id]; r)$
return σ

$\underline{\mathsf{STCRCSIG}_{(pp,pk_A,\mathbf{sk_S},\mathsf{L},\mathsf{corL},\mathsf{crcL},\mathsf{trans},\mathsf{trans_{fake}})}(id, m, r)}$
if $id \notin \mathsf{L} \setminus \mathsf{corL}$ **return** $\perp$
$\mathsf{crcL} \leftarrow \mathsf{crcL} \cup \{id\}$
if $b = 0$ $(\sigma, \mathsf{trans}[id]) \leftarrow \mathsf{Sign}(pp, \mathbf{sk_S}[id], pk_A, m, \mathsf{trans}[id]; r)$
if $b = 1$ $(\sigma, \mathsf{trans_{fake}}[id]) \leftarrow \mathsf{FakeSign}(pp, \mathbf{sk_S}[id], pk_A, m, \mathsf{trans}[id], \mathsf{trans_{fake}}[id]; r)$
return σ

Fig. 2: Oracles used in the security experiments defined in Figure 3.

$\underline{\mathsf{Exp}^{\mathsf{IND1},b}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)}$
$\mathbf{pk_S}, \mathbf{sk_S}, \mathsf{trans}, \mathsf{trans_{fake}} \leftarrow ()$
$\mathsf{L}, \mathsf{crcL}, \mathsf{corL}, \mathsf{Query} \leftarrow \emptyset$
$pp \leftarrow \mathsf{Setup}(1^\lambda)$
$(pk_A, sk_A) \leftarrow \mathsf{AKGen}(pp)$
$st \leftarrow \mathcal{A}_1^{\mathsf{ADDU},\mathsf{CRPT},\mathsf{SIGN},\mathsf{CRCSIG}}(pp, pk_A)$
$b' \leftarrow \mathcal{A}_2^{\mathsf{COERCE}}(st)$
return b'

$\underline{\mathsf{Exp}^{\mathsf{IND2},b}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)}$
$\mathbf{pk_S}, \mathbf{sk_S}, \mathsf{trans}, \mathsf{trans_{fake}} \leftarrow ()$
$\mathsf{L}, \mathsf{crcL}, \mathsf{corL}, \mathsf{Query} \leftarrow \emptyset$
$pp \leftarrow \mathsf{Setup}(1^\lambda)$
$(pk_A, sk_A) \leftarrow \mathsf{AKGen}(pp)$
$b' \leftarrow \mathcal{A}^{\mathsf{ADDU},\mathsf{CRPT},\mathsf{COERCE},\mathsf{SIGN},\mathsf{CRCSIG}}(pp, pk_A)$
return b'

$\underline{\mathsf{Exp}^{\mathsf{IND3},b}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)}$
$\mathbf{pk_S}, \mathbf{sk_S}, \mathsf{trans}, \mathsf{trans_{fake}} \leftarrow ()$
$\mathsf{L}, \mathsf{crcL}, \mathsf{corL}, \mathsf{Query} \leftarrow \emptyset$
$pp \leftarrow \mathsf{Setup}(1^\lambda)$
$(pk_A, sk_A) \leftarrow \mathsf{AKGen}(pp)$
$b' \leftarrow \mathcal{A}^{\mathsf{ADDU},\mathsf{CRPT},\mathsf{COERCE},\mathsf{SIGN},\mathsf{STCRCSIG}}(pp, pk_A)$
return b'

$\underline{\mathsf{Exp}^{\mathsf{sound}}_{\mathsf{INC\text{-}SIG}}(\lambda)}$
$pp \leftarrow \mathsf{Setup}(1^\lambda)$
$(pk_A, sk_A) \leftarrow \mathsf{AKGen}(pp)$
$(pk_S, sk_S, \mathsf{trans}) \leftarrow \mathsf{SKGen}(pp, pk_A)$
$\mathsf{trans_{fake}} \leftarrow \mathsf{FakeTrans}(pp, pk_A, \mathsf{trans}, \perp)$
$(\sigma, \mathsf{trans_{fake}}) \leftarrow \mathsf{FakeSign}(pp, sk_S, pk_A, m, \mathsf{trans}, \mathsf{trans_{fake}})$
$b \leftarrow \mathsf{Authenticate}(pp, pk_S, sk_A, m, \sigma)$
return b

$\underline{\mathsf{Exp}^{\mathsf{st\text{-}sound}}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)}$
$\mathbf{pk_S}, \mathbf{sk_S}, \mathsf{trans}, \mathsf{trans_{fake}} \leftarrow ()$
$\mathsf{L}, \mathsf{crcL}, \mathsf{corL}, \mathsf{Query} \leftarrow \emptyset$
$pp \leftarrow \mathsf{Setup}(1^\lambda)$
$(pk_A, sk_A) \leftarrow \mathsf{AKGen}(pp)$
$(id^*, m^*, \sigma^*) \leftarrow \mathcal{A}^{\mathsf{ADDU},\mathsf{CRPT},\mathsf{FAKECOERCE},\mathsf{SIGN},\mathsf{FAKESIGN}}(pp, pk_A)$
if $id^* \in \mathsf{crcL} \wedge (id^*, m^*) \notin \mathsf{Query} \wedge \mathsf{Authenticate}(pp, \mathbf{pk_S}[id^*], sk_A, m^*, \sigma^*) = 1$
return 1
else **return** 0

Fig. 3: Experiments for weak receipt-freeness, strong receipt-freeness and coercion-resistance where the adversary has access to oracles defined in Figure 2.

- ***Indistinguishability*** $(\mathsf{IND3})$*: for any PPT adversary* $\mathcal{A}$*, there exists a negligible function* negl *such that*

$$\Big|\Pr\Big[\mathsf{Exp}^{\mathsf{IND3},0}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda) = 1\Big] - \Pr\Big[\mathsf{Exp}^{\mathsf{IND3},1}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda) = 1\Big]\Big| \leq \mathsf{negl}(\lambda).$$

- ***Strong soundness****: for any PPT adversary* $\mathcal{A}$*, there exists a negligible function* negl *such that*

$$\Pr\Big[\mathsf{Exp}^{\mathsf{st\text{-}sound}}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda) = 1\Big] \leq \mathsf{negl}(\lambda).$$

where $\mathsf{Exp}^{\mathsf{IND3},b}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)$ *for* $b \in \{0, 1\}$ *and* $\mathsf{Exp}^{\mathsf{st\text{-}sound}}_{\mathcal{A},\mathsf{INC\text{-}SIG}}(\lambda)$ *are the experiments defined in Figure 3.*

B. On Coercion of Signers During Key Generation

Our security model for incoercible signatures incorporates a hierarchy of incoercibility properties that are influenced and inspired by the existing literature [16], [17], [27], [28]. In particular, it captures the weakest variant of incoercibility that is described in the literature. We now show that our notion of coercion-resistance is the strongest possible form of attainable incoercibility for digital signatures.

Recall that coercion-resistance captures an attacker that can view the transcript of a coerced signer at any point *after* key generation and act on their behalf thereafter. We assume that key generation is always performed honestly, regardless of whether the signer is corrupt, coerced, or honest. We now demonstrate the following result: a definition of incoercibility in which the adversary can coerce and interact with signers during key generation is not satisfiable in the public-key setting. We consider a natural extension to coercion-resistance, that we call *strong* coercion-resistance, in which the attacker can interact with the signer during key generation. In such a model, we require a new algorithm FakeSKGen, defined as follows.

FakeSKGen(pp, pk_A, r) On input of public parameters pp, authenticator public key pk_A and randomness r provided by the attacker, FakeSKGen outputs a signer key pair (pk_S, sk_S), where pk_S is the signer's public key and sk_S is the signer's private key, and the fake transcript $\mathsf{trans}_{\mathsf{fake}}$.

We also require modifications to the ADDU oracle for both the IND3 indistinguishability and strong soundness experiments, which we detail in Figure 4. More specifically, algorithm FakeSKGen allows a signer to generate their own key pair, potentially by deviating from the honest key generation algorithm. In the indistinguishability experiment, oracle ADDU returns the output of real key generation algorithm SKGen or the output of algorithm FakeSKGen, depending on a bit b chosen in the experiment. In the strong soundness experiment, oracle ADDU returns the output of algorithm FakeSKGen for coerced signers. With the above modifications, our strong coercion-resistance definition captures an attacker that interacts with signers during key generation. We obtain the result in Lemma 1.

Lemma 1. *No construction for an incoercible signature scheme can satisfy* strong coercion-resistance*.*

Informally, this result holds because an adversary in the strong coercion-resistance experiments can always succeed. In particular, if the adversary can choose the randomness used in key generation, then they can use this randomness to generate their own public and secret key from honest key generation. The public key of the signer must match the public key held by the adversary, otherwise the adversary can break indistinguishability. Moreover, due to completeness, the attacker can use this key pair to construct authentic signatures in the strong soundness experiment. We now present a proof of Lemma 1.

Proof. First, we show that, if the public key output by $\mathsf{SKGen}(pp, pk_A; r)$ and $\mathsf{FakeSKGen}(pp, pk_A, r)$ are different with non–negligible probability ϵ, we can build an adversary $\mathcal{A}$ that wins in the indistinguishability game with non–negligible probability ϵ. $\mathcal{A}$ runs $\mathsf{SKGen}(pp, pk_A; r)$ and then simply guesses $b = 0$ if they receive the same public key and $b = 1$ if they do not receive the same public key.

Using the above, we now show that we can build an adversary $\mathcal{A}'$ that wins in the strong soundness game. Firstly, $\mathcal{A}$ selects some randomness r identically to in SKGen. They compute $(pk^*, sk^*, \mathsf{trans}) \leftarrow \mathsf{SKGen}(pp, pk_A; r)$. For any id^*, they input (id^*, r) to the ADDU oracle. They abort if the ADDU oracle outputs a different public key to pk^*, which occurs with negligible probability. Otherwise, they compute $(\sigma^*, \mathsf{trans}) \leftarrow \mathsf{Sign}(pp, sk^*, pk_A, m^*, \mathsf{trans})$ for any message m^*. Finally, they output (id^*, m^*, σ^*). Clearly, $id^* \in \mathsf{crcL}$ and the signing oracle has not been used. As the key generation and signing were performed honestly, $\mathsf{Authenticate}(pp, pk^*, sk_A, m^*, \sigma^*) = 0$ with negligible probability due to completeness. As pk^* is identical to $\mathbf{pk_S}[id^*]$, the adversary $\mathcal{A}'$ wins with non–negligible probability. □

We note that, if the signer obtains a secret input from the authenticator during key generation, the attacker is unable to choose all of the randomness for key generation. Consequently, the result in Lemma 1 can be overcome. This was indeed the approach taken by [10]. In this paper, we choose to focus on a non-interactive setting and therefore avoid the requirement of some secret input from the authenticator, which might be difficult to implement in practice.

III. Constructions

In this section, we provide two constructions of incoercible signature schemes. Our constructions employ a standard signature scheme $\mathsf{SIG} = (\mathsf{SIG.Setup}, \mathsf{SIG.KGen}, \mathsf{SIG.Sign}, \mathsf{SIG.Vf})$, which satisfies standard notions of correctness and unforgeability (EUF-CMA), and a sender-deniable encryption scheme $\mathsf{DEN} = (\mathsf{DEN.Setup}, \mathsf{DEN.KGen}, \mathsf{DEN.Enc}, \mathsf{DEN.Dec}, \mathsf{DEN.Exp})$, which, throughout this work, we refer to as a deniable encryption scheme for brevity. A deniable encryption scheme is a standard public-key encryption scheme, equipped with an additional algorithm DEN.Exp that, for a ciphertext c that encrypts message m and is output by encryption algorithm DEN.Enc, generates randomness such that c appears to encrypt an alternative message m'. We require that the deniable encryption scheme satisfies correctness and indistinguishability under chosen plaintext attacks (IND-CPA), as in a traditional encryption scheme, and indistinguishability of explanations (IND-EXP), which ensures that the randomness output by algorithm DEN.Exp is indistinguishable from the real randomness used by algorithm DEN.Enc. We recall the syntax and security models for these building blocks in [29].

A. A Coercion-Resistant Construction

We introduce a construction that we call CR.SIG that satisfies our strongest form of incoercibility: coercion-resistance. We present CR.SIG in Figure 5: it relies on a standard signature scheme SIG and a deniable encryption scheme DEN. Additionally, our construction uses two hash functions $\mathcal{H}_1 : \{0,1\}^* \rightarrow \{0,1\}^*$, $\mathcal{H}_2 : \{0,1\}^* \rightarrow \mathcal{M}$, where $\mathcal{M}$ is the message space of the signature scheme SIG. The first is assumed to model a random oracle, whereas the second is required to be collision resistant.

Our construction works as follows. The authenticator generates a key pair for a deniable encryption scheme. During key generation, in addition to generating a key pair for signature

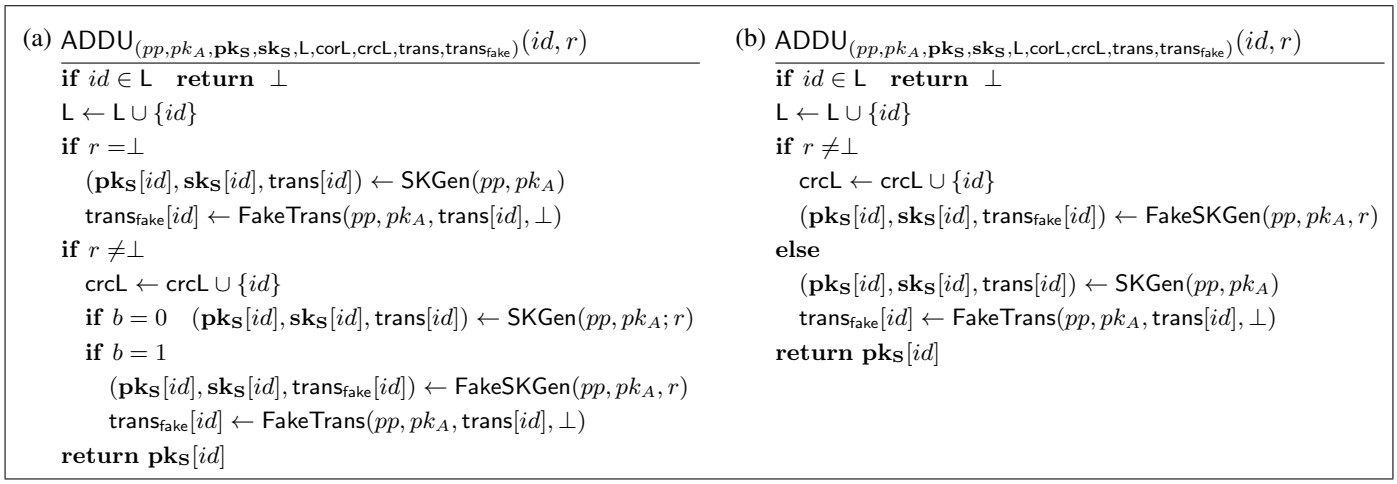

Fig. 4: The modified ADDU oracles for the (a) indistinguishability and (b) strong soundness experiments.

scheme SIG, a signer generates a random string s and deniably encrypts this under the authenticator's public key. The signer's secret key consists of a secret key for SIG and a string s. The corresponding public key consists of the public key for SIG and the ciphertext that encrypts s. A signature consists of a deniable encryption of s, as well as a standard signature that signs both the message and the deniable encryption ciphertext. The authenticator can detect coercion by decrypting the ciphertexts contained in the public key and the signature, and comparing the two, via the Authenticate algorithm. The signer creates a fake transcript that indicates s', rather than s, is contained in the signer's secret key. In this way, by security of the deniable encryption scheme, the attacker cannot distinguish a real and a fake transcript. Moreover, the coercive attacker cannot forge an authentic signature without knowledge of s, and our construction achieves strong soundness.

Our coercion-resistance construction satisfies correctness, completeness, unforgeability and coercion-resistance, as defined in Section II. We obtain the formal result in Theorem 1, which we prove in [29]. Here, we provide a proof sketch of the result.

Theorem 1. *Let* SIG *and* DEN *be a secure signature scheme and deniable encryption scheme respectively, as defined in [29], and the hash functions $\mathcal{H}_1$ and $\mathcal{H}_2$ be modelled as a random oracle model and collision resistant respectively. Then,* CR.SIG *is a secure construction of a coercion-resistant incoercible signature scheme. That is,* CR.SIG *safisfies correctness, completeness, unforgeability and coercion-resistance.*

Proof sketch. Trivially, correctness and completeness of CR.SIG follow from correctness of the building blocks used in our construction. Unforgeability follows from the EUF-CMA security of SIG and the fact that hash function $\mathcal{H}_2$ is collision resistant. Indeed, our proof of unforgeability demonstrates that, if we assume that CR.SIG does not satisfy unforgeability and $\mathcal{H}_2$ is collision resistant, then it is possible to construct an adversary that succeeds in breaking the EUF-CMA security of SIG. Then, by contradiction, the result holds.

To prove IND3 indistinguishability, we proceed through a series of game hops that we show are indistinguishable to the adversary. In the final game, the view of the adversary is identical for $b = 0$ and $b = 1$. In Game 1, the experiment only generates a fake transcript once a signer is added to the list of coerced signers. As this change is superficial, i.e., the fake transcript is only required for coerced signers, this game hop is indistinguishable to the adversary. In Game 2, we use the real secret s in the fake transcript, rather than the fake secret s'. In Game 3, we also include the real encryption randomness in the fake transcript, instead of using DEN.Exp to produce fake randomness. The hops from Game 1 to Game 2 and Game 2 to Game 3 require a hybrid argument such that the reduction operates in k_1 steps, where k_1 is the number of queries made by the adversary to oracle ADDU. Ultimately, the hop from Game 1 to 2 is indistinguishable due to the IND-CPA security of the deniable encryption scheme and the hop from Game 2 to 3 is indistinguishable due to the IND-EXP security of the deniable encryption scheme. Finally, in Game 3, the fake transcript is identical to the real transcript. Moreover, all signatures generated during the experiment are the output of the real signing algorithm. Therefore, the inputs to the adversary are independent of b and indistinguishability holds.

Finally, we prove that CR.SIG satisfies strong soundness in the random oracle model. We show that, if there exists an adversary that succeeds in the strong soundness experiment, then it is possible to construct an adversary that breaks the IND-CPA property of deniable encryption scheme DEN, if it is assumed that hash function $\mathcal{H}_1$ is a random oracle. This is because, to win, the adversary must output a signer id^* alongside an encryption of $\mathcal{H}_1(m^*||s)$, where s is the secret key of the signer id^*. Therefore, they must have input $(m^*||s)$ to the $\mathcal{H}_1$ random oracle. In our reduction we show that if they can do so we can break the IND-CPA security of the deniable encryption scheme DEN. Then, the result holds by contradiction. □

B. A Strong Receipt-Free Construction

We present a strong receipt-free incoercible signature scheme construction (RF.SIG) in Figure 6 that uses a standard signature

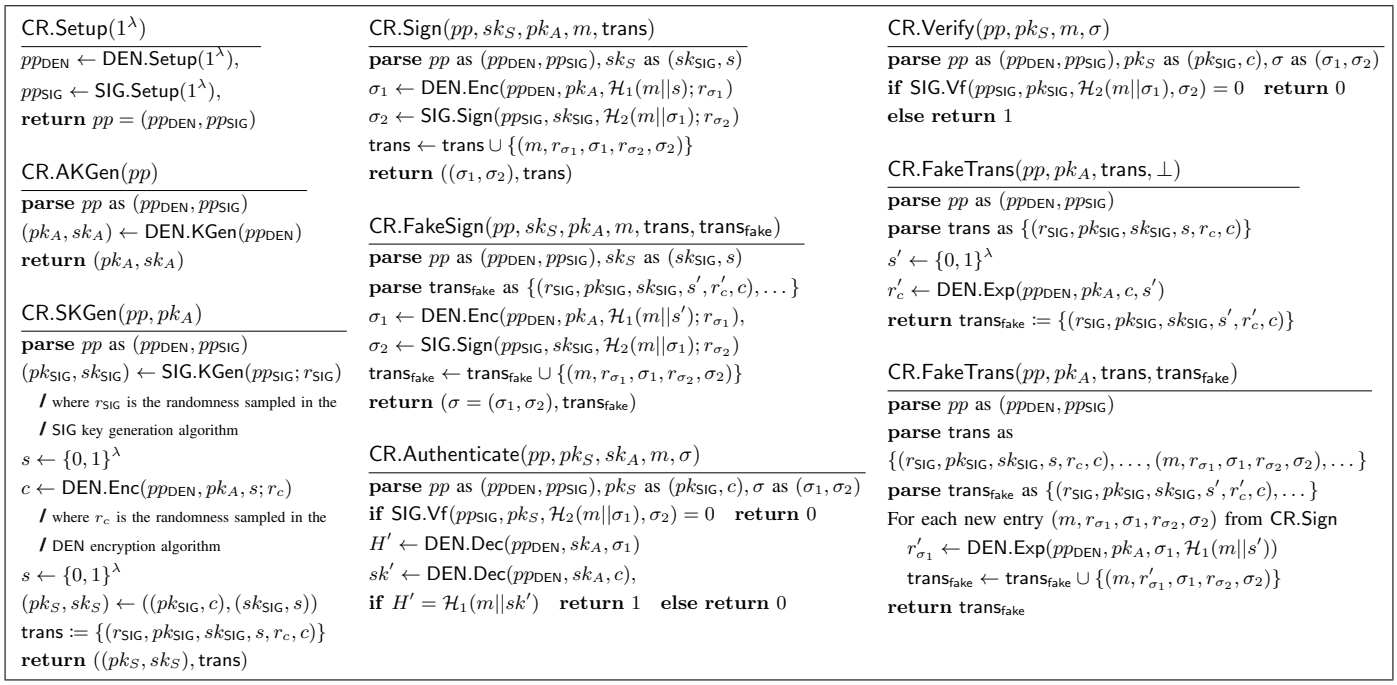

Fig. 5: Our coercion-resistant construction CR.SIG.

scheme SIG and a deniable encryption scheme DEN as building blocks. We also use a collision resistant hash function $\mathcal{H} : \{0,1\}^* \rightarrow \mathcal{M}$, where $\mathcal{M}$ is the message space of the signature scheme SIG.

We briefly note the key differences between our constructions. During key generation, a signer only generates a key pair for signature scheme SIG, and no longer generates a random string s. To sign a message, the signer now generates an encryption of a bit that denotes whether the signer is being coerced, i.e., the signer encrypts 1 to indicate a genuine signature and 0 to indicate coercion. Signing then proceeds as in our coercion-resistant construction. The authenticator can decrypt this bit (and consequently detect coercion) via the Authenticate algorithm. By using deniable encryption, our construction ensures that, when a signer is coerced, they can produce fake randomness such that they appear to have encrypted a different bit.

Our strong receipt-freeness construction RF.SIG does not satisfy coercion-resistance. In fact, if the attacker can obtain a coerced signer's transcript and provide a message and randomness to the signer, the coerced signer cannot output a fake transcript that will convince the attacker that the signer cooperated. Hence, RF.SIG cannot satisfy the IND3 indistinguishability requirement of coercion-resistance. Moreover, the fake transcript of a coerced signer contains the real secret key of the signer, which a coercive attacker can use to create valid and authentic forgeries on behalf of the signer. As such, RF.SIG cannot satisfy the strong soundness property that is necessary for coercion-resistance.

We show that our strong receipt-free construction satisfies correctness, completeness, unforgeability and strong receipt-freeness, as defined in Section II. In fact, we obtain Theorem 2, which we prove formally in [29], and for which we provide a proof sketch here.

Theorem 2. *Let* SIG *and* DEN *be a secure signature scheme and deniable encryption scheme respectively, as defined in [29], and the hash function $\mathcal{H}$ be collision resistant. Then,* RF.SIG *is a secure construction of a strong receipt-free incoercible signature scheme. That is,* RF.SIG *satisfies correctness, completeness, unforgeability and strong receipt-freeness.*

Proof sketch. Correctness, completeness and soundness (required for strong receipt-freeness) of RF.SIG follow trivially from correctness of the signature scheme SIG and the deniable encryption scheme DEN. The proof of unforgeability is very similar to the unforgeability proof of our coercion-resistant construction. That is, it follows from the EUF-CMA security of SIG and the fact that hash function $\mathcal{H}$ is collision resistant.

This leaves us to show that our construction satisfies the IND2 indistinguishability requirement of strong receipt-freeness. Indistinguishability holds as a result of the IND-CPA and IND-EXP properties of the deniable encryption scheme. To prove indistinguishability, we proceed through a series of game hops, demonstrating that the hops are indistinguishable to the adversary. In our first game hop, if $b = 1$, we change oracle CRCSIG to encrypt 1 rather than 0 when generating the fake signature. In our second game hop, if $b = 1$, we attach the real randomness used to encrypt to the fake transcript, rather than the randomness generated via algorithm DEN.Exp. These hops are indistinguishable if the deniable encryption scheme satisfies IND-CPA and IND-EXP security respectively. Through these game hops we arrive at a game in which the view of the adversary is identical for $b = 0$ and $b = 1$. In particular, regardless of bit b, the adversary views a signature

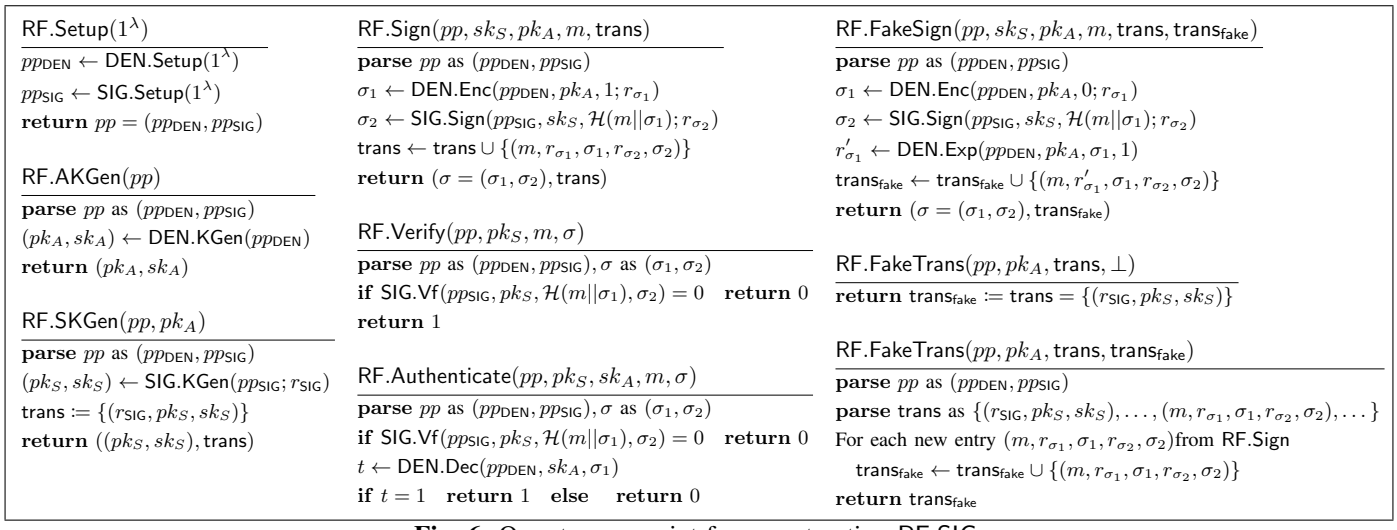

Fig. 6: Our strong receipt-free construction RF.SIG.

that contains an encryption of a bit 1 and views a transcript that contains the real encryption randomness. □

C. *Efficiency of* RF.SIG *and* CR.SIG

The efficiency of our constructions is determined by the efficiency of the sender-deniable encryption scheme. Since deniable encryption was first introduced in [22], several deniable encryption constructions have been presented in the literature, for example, [21], [23]–[25], [31]–[33]. Many constructions have improved upon the efficiency of Canetti *et al.*'s constructions [22]. Specifically, [25] proposes an efficient construction of deniable encryption based on indistinguishable obfuscation, and, more recently, an efficient construction has been proposed in the quantum setting [33]. We are hopeful that further advancement in this space can lead to even more efficient incoercible signature schemes.

RF.SIG and CR.SIG use deniable encryption in distinct ways, which leads to efficiency differences. We now provide a brief efficiency comparison of our constructions. With respect to authenticator key generation and public verification, the efficiency of both constructions is identical. Certainly, the efficiency corresponds to the key generation for the deniable encryption scheme and public verification of the underlying signature scheme, respectively. However, our coercion-resistant construction requires additional computation for signer key generation, signing and authentication. In our strong receipt-free construction, the efficiency of the signer's key generation maps directly to the efficiency of key generation for the underlying signature scheme. Our coercion-resistant construction additionally requires the computation of a deniable encryption of a string that is included in the signer's public key. Furthermore, during signing, both constructions require the computation of a single deniable encryption and a signature. However, our strong receipt-free construction only requires an encryption of a single bit, but our coercion-resistant construction requires the encryption of a string. Finally, to authenticate, our strong receipt-free construction requires the decryption of a single bit, whereas our coercion-resistant construction requires the decryption of two ciphertexts that each encrypt a string. Therefore, it is clear that our strong receipt-free construction is more efficient, though this comes at the cost of a weaker notion of security.

D. *Related Constructions*

Our incoercible signature scheme constructions are closely related to the constructions of embedded secret signature scheme constructions presented in [10], [11]. Here, we present a brief comparison of these constructions.

In [10], an embedded secret signature scheme construction is presented that is similar to our coercion-resistant construction, and provides identical efficiency in terms of the sizes of signatures and computation during signing, verification and authentication. Indeed, the construction in [10] and both our receipt-free and coercion-resistant constructions use deniable encryption and require that a secret is shared between the authenticator and signer. Nevertheless, our constructions can be distinguished with respect to how the secret is transmitted. The construction in [10] assumes that the authenticator and signer can privately share a secret during key generation. We do not require such an assumption and, indeed, our syntax models key generation as non-interactive. Instead, our constructions allow the signer to generate a secret during key generation (coercion-resistant construction) or simply encrypt a bit during signing (receipt-free construction).

Furthermore, our contributions differ from those in [10] with respect to the security models. Overall, our security model has a similar approach to that of embedded secret signature schemes [10]. In fact, embedded secret signatures must satisfy an indistinguishability requirement and a soundness requirement. However, we distinguish our security model in the following ways. Firstly, our security model captures a spectrum of incoercibility notions that reflects the understanding of incoercibility established in the literature. Secondly, our indistinguishability notions are stronger. More specifically,

the security model in [10] captures an indistinguishability experiment similar to our IND1 indistinguishability property. However, the attacker is assumed to only access a real or fake secret key of a coerced signer, rather than a full transcript. By contrast, in our security model, the adversary is given the signer's full transcript. Finally, we highlight that the notion of soundness introduced in [10], called embedded secret unforgeability, ensures that an attacker with a fake secret key cannot output a signature without an embedded warning. Our notion of coercion-resistance captures a similar soundness property, in addition to a stronger indistinguishability property.

We note that, in [11], a very efficient construction for an embedded secret signature scheme is also given. In fact, this construction is more efficient than both our strong receipt-free and coercion-resistant constructions. However, it does not come with an accompanying security model, and, in fact, does not consider an attacker that demands a signer's secret key. Our constructions, on the other hand, are accompanied with rigorous proofs under suitable security definitions.

IV. On Incoercibility and Deniability

Deniability is closely linked to coercion. In fact, both our strong receipt-free and coercion-resistant constructions make use of deniable encryption. This raises the question: how are deniable encryption and incoercible signatures related? In this section, we answer the question by showing that given a weak receipt-free incoercible signature scheme we can build a *partial* deniable encryption scheme. First, we provide a formal definition for a partial deniable encryption scheme. This is a deniable encryption scheme that only encrypts a single bit, i.e., the message space is $\{0,1\}$, and can only explain one of two messages, e.g., the message $m = 0$. That is, given a ciphertext c, $\mathsf{DEN.Exp}$ can only generate randomness such that c appears to encrypt 0, regardless of the message it encrypts. We then show that a secure partial deniable encryption construction can be built from a weak receipt-free incoercible signature scheme. Therefore, we formally show that any construction of a weak receipt-free incoercible signature scheme will either make use of partial deniable encryption as a building block, or (if more efficient than a construction using partial deniable encryption) lead to efficiency improvements for partial deniable encryption. We leave as an open question whether partial deniable encryption schemes can be built more efficiently than standard deniable encryption schemes, leading to efficiency improvements for weak receipt-free constructions.

A. Partial Deniable Encryption

We adapt the definition of public-key sender-deniable encryption [22], [25] such that the message space is $\{0,1\}$ and the explanation algorithm no longer takes as input a message, because the only message that can be explained is 0. Additionally, we modify definitions of correctness, IND-CPA and IND-EXP to the partial deniability setting. In particular, the IND-CPA experiment does not require that the adversary output two messages because the only possible messages are 0 and 1. In the IND-EXP experiment, as only the message 0 can be explained, the adversary does not output a message.

Definition 8 (Partial Deniable Encryption Scheme)**.** *A* partial deniable encryption scheme (PDEN) *is a tuple of PPT algorithms* (PDEN.Setup, PDEN.KGen, PDEN.Enc, PDEN.Dec, PDEN.Exp) *such that:*

- $\mathsf{PDEN.Setup}(1^\lambda)$ On input of security parameter 1^λ, PDEN.Setup outputs public parameters pp_{PDEN}.
- $\mathsf{PDEN.KGen}(pp_{\mathsf{PDEN}})$ On input of public parameters pp_{PDEN}, PDEN.KGen outputs a key pair $(pk_{\mathsf{PDEN}}, sk_{\mathsf{PDEN}})$ where pk_{PDEN} is the public encryption key and sk_{PDEN} is the private decryption key.
- $\mathsf{PDEN.Enc}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, m)$ On input of public parameters pp_{PDEN}, public key pk_{PDEN} and message $m \in \{0,1\}$, PDEN.Enc outputs a ciphertext c.
- $\mathsf{PDEN.Dec}(pp_{\mathsf{PDEN}}, sk_{\mathsf{PDEN}}, c)$ On input of public parameters pp_{PDEN}, private key sk_{PDEN} and ciphertext c, PDEN.Dec outputs a message m.
- $\mathsf{PDEN.Exp}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, c)$ On input of public parameters pp_{PDEN}, public key pk_{PDEN}, and ciphertext c, PDEN.Exp outputs a string u.

Definition 9 (Correctness)**.** *A partial deniable encryption scheme* PDEN *satisfies* correctness *if, for any message* $m \in \{0,1\}$*, there exists a negligible function* negl *such that*

$$\Pr\left[\begin{array}{l} pp_{\mathsf{PDEN}} \leftarrow \mathsf{PDEN.Setup}(1^\lambda); \\ (pk_{\mathsf{PDEN}}, sk_{\mathsf{PDEN}}) \leftarrow \mathsf{PDEN.KGen}(pp_{\mathsf{PDEN}}); \\ c \leftarrow \mathsf{PDEN.Enc}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, m) \end{array} : \begin{array}{c} \mathsf{PDEN.Dec}(pp_{\mathsf{PDEN}}, \\ sk_{\mathsf{PDEN}}, c) = m \end{array}\right] \geq 1 - \mathsf{negl}(\lambda).$$

Definition 10 (IND-CPA)**.** *A partial deniable encryption scheme* PDEN *satisfies* indistinguishability under a chosen plaintext attack (IND-CPA) *if, for any PPT adversary* $\mathcal{A}$*, there exists a negligible function* negl *such that*

$$\left|\Pr\left[\mathsf{Exp}^{\mathsf{IND\text{-}CPA},0}_{\mathcal{A},\mathsf{PDEN}}(\lambda) = 1\right] - \Pr\left[\mathsf{Exp}^{\mathsf{IND\text{-}CPA},1}_{\mathcal{A},\mathsf{PDEN}}(\lambda) = 1\right]\right| \leq \mathsf{negl}(\lambda)$$

where $\mathsf{Exp}^{\mathsf{IND\text{-}CPA},b}_{\mathcal{A},\mathsf{PDEN}}(\lambda)$ *is the experiment defined in Figure 7 for* $b \in \{0,1\}$.

Definition 11 (IND-EXP)**.** *A partial deniable encryption scheme* PDEN *satisfies* indistinguishability of explanation (IND-EXP) *if, for any PPT adversary* $\mathcal{A}$*, there exists a negligible function* negl *such that*

$$\left|\Pr\left[\mathsf{Exp}^{\mathsf{IND\text{-}EXP},0}_{\mathcal{A},\mathsf{PDEN}}(\lambda) = 1\right] - \Pr\left[\mathsf{Exp}^{\mathsf{IND\text{-}EXP},1}_{\mathcal{A},\mathsf{PDEN}}(\lambda) = 1\right]\right| \leq \mathsf{negl}(\lambda)$$

where $\mathsf{Exp}^{\mathsf{IND\text{-}EXP},b}_{\mathcal{A},\mathsf{PDEN}}(\lambda)$ *is the experiment defined in Figure 7 for* $b \in \{0,1\}$.

$\mathsf{Exp}^{\mathsf{IND\text{-}CPA},b}_{\mathcal{A},\mathsf{PDEN}}(\lambda)$

$pp_{\mathsf{PDEN}} \leftarrow \mathsf{PDEN.Setup}(1^\lambda)$
$(pk_{\mathsf{PDEN}}, sk_{\mathsf{PDEN}}) \leftarrow \mathsf{PDEN.KGen}(pp_{\mathsf{PDEN}})$
$c \leftarrow \mathsf{PDEN.Enc}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, b)$
$b' \leftarrow \mathcal{A}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, c)$
return b'

$\mathsf{Exp}^{\mathsf{IND\text{-}EXP},b}_{\mathcal{A},\mathsf{PDEN}}(\lambda)$

$pp_{\mathsf{PDEN}} \leftarrow \mathsf{PDEN.Setup}(1^\lambda)$
$(pk_{\mathsf{PDEN}}, sk_{\mathsf{PDEN}}) \leftarrow \mathsf{PDEN.KGen}(pp_{\mathsf{PDEN}})$
$c \leftarrow \mathsf{PDEN.Enc}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, 0; u_0)$;
$u_1 \leftarrow \mathsf{PDEN.Exp}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, c)$;
$b' \leftarrow \mathcal{A}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, c, u_b)$
return b'

Fig. 7: Experiments for indistinguishability under a chosen plaintext attack and indistinguishability of explanation for a partial deniable encryption scheme.

B. Constructing a Partial Deniable Encryption Scheme

We show that given an incoercible signature scheme INC-SIG that satisfies weak receipt-freeness, we can build a secure partial deniable encryption scheme PDEN. That is, we show that PDEN (Figure 8) satisfies correctness, IND-CPA and IND-EXP security, as defined in section IV-A, if INC-SIG is a weak receipt-free incoercible signature scheme.

Intuitively, our result holds because a "real" signature can be used in an encryption of 0, and a "fake" signature can be used in an encryption of 1. The authenticator can differentiate between such signatures and so decrypt the ciphertext. Then, when explaining a ciphertext that encrypts 0 (resp., 1) and hence contains a real (resp., fake) signature, the real (resp., fake) transcript can be output. Only a *partial* decryption scheme can be built from incoercible signatures because an incoercible signature does not allow a transcript to be generated such that a ciphertext encrypting 0 and containing a real signature can be explained for message $m = 1$, as if it contains a fake signature.

a) Correctness.: We first show that for $m = 0$, decryption will always return 0. The ciphertext c that encrypts 0 is distributed as follows: $(pk_S, sk_S, \mathsf{trans}) \leftarrow \mathsf{INC\text{-}SIG.SKGen}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}; r)$; $(\sigma, \mathsf{trans}) \leftarrow \mathsf{INC\text{-}SIG.Sign}(pp_{\mathsf{PDEN}}, sk_S, pk_{\mathsf{PDEN}}, 0; r')$; $c \leftarrow \mathsf{trans}$. By completeness of INC-SIG, $\mathsf{INC\text{-}SIG.Authenticate}(pp_{\mathsf{PDEN}}, pk_S, sk_{\mathsf{PDEN}}, 0, \sigma) = 0$ with at most negligible probability $\mathsf{negl}(\lambda)$. Therefore, the ciphertext decrypts to 0 with probability at least $1 - \mathsf{negl}(\lambda)$.

We next show that for $m = 1$, decryption will always return 1. The ciphertext c that encrypts 1 is distributed as follows: $(pk_S, sk_S, \mathsf{trans}) \leftarrow \mathsf{INC\text{-}SIG.SKGen}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}; r)$; $\mathsf{trans}_{\mathsf{fake}} \leftarrow \mathsf{FakeTrans}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, \mathsf{trans}, \perp)$; $\sigma, \mathsf{trans}_{\mathsf{fake}}) \leftarrow \mathsf{INC\text{-}SIG.FakeSign}(pp_{\mathsf{PDEN}}, sk_S, pk_{\mathsf{PDEN}}, 0, \mathsf{trans}, \mathsf{trans}_{\mathsf{fake}})$; $c \leftarrow \mathsf{trans}_{\mathsf{fake}}$. By soundness of INC-SIG, $\mathsf{INC\text{-}SIG.Authenticate}(pp_{\mathsf{PDEN}}, pk_S, sk_{\mathsf{PDEN}}, 0, \sigma) = 1$ with at most negligible probability $\mathsf{negl}(\lambda)$. Therefore, the ciphertext decrypts to 1 with probability at least $1 - \mathsf{negl}(\lambda)$.

b) IND-CPA *security.:* Let $\mathcal{A}$ be an adversary in the $\mathsf{Exp}^{\mathsf{IND\text{-}CPA},b}_{\mathcal{A},\mathsf{PDEN}}(\lambda)$ experiment that is successful with non-negligible probability. We show that we can construct an adversary $\mathcal{A}' = (\mathcal{A}'_1, \mathcal{A}'_2)$ that succeeds in the $\mathsf{Exp}^{\mathsf{IND1},b}_{\mathcal{A}',\mathsf{INC\text{-}SIG}}(\lambda)$ experiment with non-negligible probability. We present $\mathcal{A}'$ in Figure 9. It is clear that inputs to $\mathcal{A}$ are distributed identically to the $\mathsf{Exp}^{\mathsf{IND\text{-}CPA},b}_{\mathcal{A},\mathsf{PDEN}}(\lambda)$ experiment for the bit b chosen in the IND1 indistinguishability experiment. When $b = 0$, trans is distributed identically to the encryption of 0 for the partial deniable encryption scheme PDEN. When $b = 1$, trans is distributed identically to the encryption of 1 in the partial deniable encryption scheme PDEN. Therefore, if $\mathcal{A}$ successfully guesses b in the IND-CPA experiment then $\mathcal{A}'$ will successfully guess b in the IND1 indistinguishability experiment.

c) IND-EXP *security.:* Indistinguishability of explanation is perfectly satisfied by the PDEN construction. When 0 is encrypted, r, r' is the only randomness chosen and is output in the ciphertext. PDEN.Exp returns this randomness, and inputs to the adversary are independent of b in the IND-EXP security experiment.

V. Incoercible Strong Designated Verifier Signatures

Designated verifier signature schemes [34] allow for signing with respect to a designated verifier. Only the designated verifier is able to verify that the signer generated the signature, and this conviction cannot be transferred to others. This is because the designated verifier is able to simulate signatures with respect to a signer due to the source hiding requirement [35] for such signatures, meaning that all signatures could have been authored by the designated verifier.

In this setting, if the designated verifier can be trusted not to simulate a signature, then that signature must have been authored by the signer. To ensure that signatures cannot be attributed to their signer, even when the designated verifier is trusted not to simulate signatures, this definition was strengthened in *strong* designated verifier signature schemes [30], [34], [36]. This primitive has the additional requirement that, to all but the designated verifier, a given signature could have been produced by any signer.

As the designated verifier's secret key is necessary to verify signatures, this setting seems well-suited to explore incoerciblity. Indeed, the designated verifier can now take on the role of the authenticator and detect coercion during verification. As signatures are not publicly verifiable, the attacker is not able to detect coercion evasion by verifying signatures. We therefore provide a security model for incoercible strong designated verifier signature schemes and a construction that provably satisfies coercion-resistance.

A. Security Model

Strong designated verifier signature schemes (SDVS) were informally discussed in [34] and the first formal definitions were provided in [36]. In [30], state-of-the-art security definitions

$\mathsf{PDEN.Setup}(1^\lambda)$
$pp_{\mathsf{PDEN}} \leftarrow \mathsf{INC\text{-}SIG.Setup}(1^\lambda)$
return pp_{PDEN}

$\mathsf{PDEN.KGen}(pp_{\mathsf{PDEN}})$
$(pk_{\mathsf{PDEN}}, sk_{\mathsf{PDEN}}) \leftarrow \mathsf{INC\text{-}SIG.AKGen}(pp_{\mathsf{PDEN}})$
return $(pk_{\mathsf{PDEN}}, sk_{\mathsf{PDEN}})$

$\mathsf{PDEN.Dec}(pp_{\mathsf{PDEN}}, sk_{\mathsf{PDEN}}, c)$
parse c as $\{r, pk_S, sk_S, 0, r', \sigma\}$
if $\mathsf{INC\text{-}SIG.Authenticate}(pp_{\mathsf{PDEN}}, pk_S, sk_{\mathsf{PDEN}}, 0, \sigma) = 1$
  return 0
else return 1

$\mathsf{PDEN.Enc}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, m)$
$(pk_S, sk_S, \mathsf{trans}) \leftarrow \mathsf{INC\text{-}SIG.SKGen}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}; r)$
if $m = 0$
  $(\sigma, \mathsf{trans}) \leftarrow \mathsf{INC\text{-}SIG.Sign}(pp_{\mathsf{PDEN}}, sk_S, pk_{\mathsf{PDEN}}, 0, \mathsf{trans}; r')$
  $c \leftarrow \mathsf{trans}$
if $m = 1$
  $\mathsf{trans}_{\mathsf{fake}} \leftarrow \mathsf{FakeTrans}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, \mathsf{trans}, \perp)$
  $(\sigma, \mathsf{trans}_{\mathsf{fake}}) \leftarrow \mathsf{INC\text{-}SIG.FakeSign}(pp_{\mathsf{PDEN}}, sk_S, pk_{\mathsf{PDEN}}, 0, \mathsf{trans}, \mathsf{trans}_{\mathsf{fake}})$
  $c \leftarrow \mathsf{trans}_{\mathsf{fake}}$
return c

$\mathsf{PDEN.Exp}(pp_{\mathsf{PDEN}}, pk_{\mathsf{PDEN}}, c)$
parse c as $\{r, pk_S, sk_S, 0, r', \sigma\}$
return (r, r')

Fig. 8: A partial deniable encryption scheme from a weak receipt-free incoercible signature scheme.

$\mathcal{A}'^{\mathsf{ADDU,CRPT,SIGN,CRCSIG}}_1(pp, pk_A)$
choose any id
$pk_S \leftarrow \mathsf{ADDU}(id)$
$\sigma \leftarrow \mathsf{CRCSIG}(id, 0)$

$\mathcal{A}'^{\mathsf{COERCE}}_2(st)$
$\mathsf{trans} \leftarrow \mathsf{COERCE}(id)$
$b' \leftarrow \mathcal{A}(pp, pk_A, \mathsf{trans})$
return b'

Fig. 9: Adversary $\mathcal{A}'$ that breaks the IND1 indistinguishability of weak receipt-free incoercible signature scheme INC-SIG given an adversary that can break the IND-CPA security of partial deniable encryption scheme PDEN.

for SDVS were provided, based on the model for designated verifier signature schemes given in [35]. We detail the security model for a strong designated verifier signature scheme in full in [29].

We base our syntax and security model for incoercible strong designated verifier schemes on [30]. Unlike standard signatures, our syntax does not need the Authenticate algorithm because the designated verifier can detect coercion during verification. This is because a secret key is now required to verify a signature and so the attacker cannot detect that a signer has evaded coercion. Incoercible SDVS must satisfy *correctness*, *unforgeability*, *source hiding*, *privacy of signer's identity*, as in standard SDVS [30]. Additionally, they must satisfy *incoercibility*, whether weak/ strong receipt-freeness or coercion-resistance. These requirements are defined similarly to those for the publicly verifiable schemes, adapted to the strong designated verifier setting. We provide a full security model for incoercible SDVS that captures these properties in [29].

a) Relation Between Incoercibility and Non–Delegatability: In [37] the non–delegatability requirement was introduced for SDVS. This ensures a signer cannot delegate their signing rights to another entity. Intuitively, this requires that if an adversary produces a valid signature for signer S and designated verifier V, then it must know either the secret key for S or for V. Although non-delegatability and coercion–resistance seem related at first glance, they address different attack models. Coercion-resistance ensures that coerced signatures are detected, without the attacker discovering that the signer evaded coercion. Non-delegatability prevents a signer that is happy to delegate their signing rights from doing so without revealing their signing key. In both cases delegation is prevented. For coercion-resistance the signer does not wish to delegate their signing rights but is being bribed or blackmailed, whereas for non–delegatability they do wish to delegate their signing rights but not reveal their secret key.

b) Relation to Deniability.: In Section I-A we discuss how strong designated verifier signature schemes satisfy elements of deniability. Indeed, in an SDVS, if a signer is being coerced by an attacker *not* to sign a message of their choice, the signer can simply sign the message anyway and claim that it was authored by a different signer or by the designated verifier. Our incoercible strong designated verifier signature schemes inherit this deniability property, as well as provide resistance to the coercive attacks captured by incoercibility, whereby an attacker instructs a signer to sign a message of their choice.

B. A Coercion-Resistant Construction

We provide a construction that satisfies our coercion-resistance security definition for incoercible strong designated verifier schemes. The full construction and security proofs are given in [29] and we provide an intuition here.

Our construction makes use of a strong designated verifier signature scheme SDVS and a deniable encryption scheme DEN as building blocks. The idea behind the construction is similar to that used in the public verification setting, but with the strong designated verifier signature scheme SDVS replacing the standard signature scheme SIG. The SDVS building block ensures the properties necessary for a strong designated verifier signature scheme still hold. Coercion-resistance is provided using a similar argument as for the publicly verifiable construction. We note that the SDVS scheme must satisfy *strong unforgeability* [38], to ensure that our construction satisfies the privacy of signer's identity requirement.

VI. Conclusion

We introduced and defined incoercible signatures, and presented an accompanying security model. Our security model captures a strong notion of incoercibility, coercion-resistance, and we contributed an incoercible signature scheme construction that provably satisfies coercion-resistance. Additionally, our security model captures strong and weak receipt-freeness. Though these are weaker notions of security than coercion-resistance, they may be sufficient in some application scenarios. For example, strong receipt-freeness is sufficient if it can be assumed that the attacker will not attempt to produce signatures on behalf of coerced signers. Moreover, our strong receipt-freeness construction is more efficient than our coercion-resistant construction, demonstrating that efficiency/security trade-offs are possible. We comment that even more efficient weak receipt-free constructions may be possible, and leave this as an open problem.

In this work, we show that our syntax and security model can be extended in an intuitive way to the designated verifier signature scheme setting. We also present a construction that satisfies our security model in this setting. An interesting area of future research is to consider incoercibility in the context of other signing and anonymity protocols, for example, group and ring signatures.

References

[1] S. Goldwasser, S. Micali, and R. L. Rivest, "A digital signature scheme secure against adaptive chosen-message attacks," *SIAM Journal on Computing*, vol. 17, no. 2, pp. 281–308, 1988.

[2] M. Bellare and S. K. Miner, "A forward-secure digital signature scheme," in *CRYPTO'99* (M. Wiener, ed.), (Berlin, Heidelberg), pp. 431–448, Springer Berlin Heidelberg, 1999.

[3] Y. Dodis, J. Katz, S. Xu, and M. Yung, "Key-insulated public key cryptosystems," in *EUROCRYPT'02* (L. R. Knudsen, ed.), (Berlin, Heidelberg), pp. 65–82, Springer Berlin Heidelberg, 2002.

[4] Y. Dodis, J. Katz, S. Xu, and M. Yung, "Strong key-insulated signature schemes," in *PKC'03* (Y. G. Desmedt, ed.), (Berlin, Heidelberg), pp. 130–144, Springer Berlin Heidelberg, 2002.

[5] J. Håstad, J. Jonsson, A. Juels, and M. Yung, "Funkspiel schemes: an alternative to conventional tamper resistance," in *CCS'00*, pp. 125–133, Association for Computing Machinery, 2000.

[6] G. Itkis, "Cryptographic tamper evidence," in *CCS'03*, CCS '03, (New York, NY, USA), pp. 355–364, Association for Computing Machinery, 2003.

[7] G. Itkis and L. Reyzin, "Sibir: Signer-base intrusion-resilient signatures," in *CRYPTO'02* (A. Boldyreva and D. Micciancio, eds.), (Cham), pp. 499–514, Springer International Publishing, 2002.

[8] G. Itkis and P. Xie, "Generalized key-evolving signature schemes or how to foil an armed adversary," in *ACNS'03* (J. Zhou, M. Yung, and Y. Han, eds.), (Berlin, Heidelberg), pp. 151–168, Springer Berlin Heidelberg, 2003.

[9] D. Naccache, D. Pointcheval, and C. Tymen, "Monotone signatures," in *Financial Cryptography* (P. Syverson, ed.), (Berlin, Heidelberg), pp. 305–318, 2002.

[10] K. Durnoga, J. Pomykała, and T. Trabszys, "Digital signature with secretly embedded warning," *Control and Cybernetics*, vol. 42, no. 4, pp. 805–824, 2013.

[11] M. Kutyłowski and P. Kubiak, "Lightweight digital signature with secretly embedded warning," *Control and Cybernetics*, vol. 42, no. 4, pp. 825–827, 2013.

[12] S. Delaune, S. Kremer, and M. Ryan, "Verifying privacy-type properties of electronic voting protocols," *Journal of Computer Security*, vol. 17, no. 4, pp. 435–487, 2009.

[13] J. Benaloh and D. Tuinstra, "Receipt-free secret-ballot elections (extended abstract)," in *STOC'94*, STOC '94, pp. 544–553, Association for Computing Machinery, 1994.

[14] A. Juels, D. Catalano, and M. Jakobsson, "Coercion-resistant electronic elections," in *WPES'05*, pp. 61–70, 2005.

[15] J. Alwen, R. Ostrovsky, H.-S. Zhou, and V. Zikas, "Incoercible multi-party computation and universally composable receipt-free voting," in *CRYPTO'15*, pp. 763–780, 2015.

[16] R. Canetti and R. Gennaro, "Incoercible multi-party computation," in *FOCS'96*, pp. 504–513, IEEE, 1996.

[17] D. Unruh and J. Müller-Quade, "Universally composable incoercibility," in *CRYPTO'10* (T. Rabin, ed.), (Berlin, Heidelberg), pp. 411–428, Springer Berlin Heidelberg, 2010.

[18] Y. Desmedt and Y. Frankel, "Threshold cryptosystems," in *CRYPTO'89* (G. Brassard, ed.), (New York, NY), pp. 307–315, Springer New York, 1990.

[19] A. Herzberg, M. Jakobsson, S. Jarecki, H. Krawczyk, and M. Yung, "Proactive public key and signature systems," in *CCS'97*, pp. 100–110, Association for Computing Machinery, 1997.

[20] R. Ostrovsky and M. Yung, "How to withstand mobile virus attacks," in *PODC'91*, pp. 51–59, Association for Computing Machinery, 1991.

[21] R. Bendlin, J. B. Nielsen, P. S. Nordholt, and C. Orlandi, "Lower and upper bounds for deniable public-key encryption," in *ASIACRYPT'11* (D. H. Lee and X. Wang, eds.), pp. 125–142, Springer Berlin Heidelberg, 2011.

[22] R. Canetti, C. Dwork, M. Naor, and R. Ostrovsky, "Deniable encryption," in *CRYPTO'97* (B. S. Kaliski, ed.), (Berlin, Heidelberg), pp. 90–104, Springer Berlin Heidelberg, 1997.

[23] R. Canetti, S. Park, and O. Poburinnaya, "Fully deniable interactive encryption," in *Advances in Cryptology – CRYPTO'20* (D. Micciancio and T. Ristenpart, eds.), (Cham), pp. 807–835, 2020.

[24] A. O'Neill, C. Peikert, and B. Waters, "Bi-deniable public-key encryption," in *Advances in Cryptology – CRYPTO'11* (P. Rogaway, ed.), (Berlin, Heidelberg), pp. 525–542, 2011.

[25] A. Sahai and B. Waters, "How to use indistinguishability obfuscation: deniable encryption, and more," in *STOC'14*, pp. 475–484, 2014.

[26] J. Katz and Y. Lindell, *Introduction to modern cryptography*. CRC press, 2014.

[27] P. Chaidos, V. Cortier, G. Fuchsbauer, and D. Galindo, "Beleniosrf: a non-interactive receipt-free electronic voting scheme," in *CCS'16–ACM SIGSAC Conference on Computer and Communications Security*, pp. 1614–1625, 2016.

[28] A. Kiayias, T. Zacharias, and B. Zhang, "End-to-end verifiable elections in the standard model," in *EUROCRYPT'15–International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 468–498, 2015.

[29] A. Fraser, L. Garms, and E. A. Quaglia, "On the incoercibility of digital signatures." Cryptology ePrint Archive, Paper 2023/054, 2023. https://eprint.iacr.org/2023/054.

[30] F. Laguillaumie and D. Vergnaud, "Designated verifier signatures: Anonymity and efficient construction from any bilinear map," in *SCN'04*, pp. 105–119, Springer, 2004.

[31] M. Dürmuth and D. M. Freeman, "Deniable encryption with negligible detection probability: An interactive construction," in *Advances in Cryptology – EUROCRYPT 2011* (K. G. Paterson, ed.), (Berlin, Heidelberg), pp. 610–626, Springer Berlin Heidelberg, 2011.

[32] S. Agrawal, S. Goldwasser, and S. Mossel, "Deniable fully homomorphic encryption from learning with errors," in *CRYPTO'21*, pp. 641–670, 2021.

[33] A. Coladangelo, S. Goldwasser, and U. Vazirani, "Deniable encryption in a quantum world," in *STOC'22*, pp. 1378–1391, 2022.

[34] M. Jakobsson, K. Sako, and R. Impagliazzo, "Designated verifier proofs and their applications," in *EUROCRYPT'96*, pp. 143–154, Springer, 1996.

[35] R. Steinfeld, H. Wang, and J. Pieprzyk, "Efficient extension of standard schnorr/rsa signatures into universal designated-verifier signatures," in *PKC'04*, pp. 86–100, Springer, 2004.

[36] S. Saeednia, S. Kremer, and O. Markowitch, "An efficient strong designated verifier signature scheme," in *ICISC'03*, pp. 40–54, Springer, 2003.

[37] H. Lipmaa, G. Wang, and F. Bao, "Designated verifier signature schemes: attacks, new security notions and a new construction," in *International Colloquium on Automata, Languages, and Programming*, pp. 459–471, 2005.

[38] H. Tian, Z. Jiang, Y. Liu, and B. Wei, "A systematic method to design strong designated verifier signature without random oracles," *Cluster computing*, vol. 16, no. 4, pp. 817–827, 2013.