A Deniability Analysis of Signal's Initial Handshake PQXDH

Rune Fiedler
Technische Universität Darmstadt
Darmstadt, Germany
rune.fiedler@cryptoplexity.de

Christian Janson
Technische Universität Darmstadt
Darmstadt, Germany
christian.janson@cryptoplexity.de

ABSTRACT

Many use messaging apps such as Signal to exercise their right to private communication. To cope with the advent of quantum computing, Signal employs a new initial handshake protocol called PQXDH for post-quantum confidentiality, yet keeps guarantees of authenticity and deniability classical. Compared to its predecessor X3DH, PQXDH includes a KEM encapsulation and a signature on the ephemeral key. In this work we show that PQXDH does not meet the same deniability guarantees as X3DH due to the signature on the ephemeral key. Our analysis relies on plaintext awareness of the KEM, which Signal's implementation of PQXDH does not provide. As for X3DH, both parties (initiator and responder) obtain different deniability guarantees due to the asymmetry of the protocol.

For our analysis of PQXDH, we introduce a new model for deniability of key exchange that allows a more fine-grained analysis. Our deniability model picks up on the ideas of prior work and facilitates new combinations of deniability notions, such as deniability against malicious adversaries in the big brother model, i.e. where the distinguisher knows all secret keys. Our model may be of independent interest.

KEYWORDS

Deniable Key Exchange, Signal, X3DH, PQXDH

1 INTRODUCTION

Nowadays, messaging applications are widely adopted: Billions of people use messaging apps such as Signal and WhatsApp.[1] In order to provide confidentiality of the users' communication, both apps offer end-to-end encrypted messaging by using the Signal protocol. The Signal protocol consists of an initial handshake, which allows two parties to derive a shared session key, and a ratcheting mechanism [42], which provides forward secrecy and post-compromise security. Until 2023, Signal used the X3DH protocol [43] for the initial handshake. To resist the threat of "harvest now decrypt later attacks" [38] posed by future quantum computers, Signal replaced X3DH with the PQXDH protocol; pursuing post-quantum *confidentiality* (without addressing quantum attackers against authenticity or deniability). Similarly, this paper does not consider deniability against quantum adversaries.

Deniability is a subtle privacy property that allows a party to plausibly deny an action. Consider the following example in a scenario without computers and smartphones present, where Alice and Bob have a conversation. Later, Alice tells Charlie she talked to Bob, and Bob tells Charlie he never talked to Alice. Charlie does not know whose word to trust; hence, Bob can *deny* his conversation with Alice since their conversation did not leave any evidence behind. For sensitive topics such as medical conditions, sexuality, or coordinating a protest against an oppressive regime, this type of privacy guarantee can be not just helpful but even vital.

Conversations in the digital world entail sending data over the internet. Moving our previous example to conversing via messaging, Alice and Bob exchange messages as specified by the messaging protocol, resulting in a *protocol transcript*. This transcript may convince Charlie that Alice and Bob *did* converse and *what* they conversed about. However, if Alice and Bob use a deniable protocol, then either party can deny involvement in a protocol run.

A deniable key exchange is necessary (but not sufficient, as pointed out in [15]) to build deniable messaging. Signal desires deniability[2] and, therefore, a deniable initial handshake, i.e. previously X3DH and now PQXDH [38, 43]. Before X3DH, Signal used a derivative of Off-the-Record Messaging (OTR) [45], which also employs a deniable key exchange.

For a key exchange protocol to be deniable, the transcript must not prove Bob's involvement in the conversation. In particular, there must exist a Fake algorithm that *simulates* Bob's involvement. Hence, Alice acting as the *adversary* can simulate a transcript on her own by using the Fake algorithm. Charlie acting as *distinguisher* must not be able to tell the simulated transcript apart from a real transcript. Honest parties would likely use the freshly established session key in some subsequent protocol. Hence, the Fake algorithm must simulate the session key indistinguishably as well.

Several deniability definitions were produced by prior works [9, 10, 17, 19, 23, 30, 32–34, 59]; each definition follows the outline in the previous paragraph. Additionally, [23] introduces *auxiliary information* (e.g. valid protocol transcripts anybody could have observed a priori) that all algorithms (adversary, their equivalent of the Fake algorithm, and the distinguisher) can access. The definitions from the literature are incomparable, vary in the precise guarantee they capture and which formalisms they use. In consequence, capturing several deniability notions may require using several formalisms. Instead, we propose a new model for deniable key exchange, which integrates cherry-picked ideas from prior works: Our model is parameterized in the auxiliary info (*Are some protocol transcripts available?*), the power of the adversary (*Does the adversary follow the protocol?*), the power of the Fake algorithm (*How much help do you need to simulate a transcript?*), the power of the distinguisher (*Is the distinguisher a* big brother *who can force everybody to give up their secret keys? Does it learn the auxiliary info?*).

[1] https://en.wikipedia.org/wiki/WhatsApp#User_statistics

[2] https://signal.org/blog/simplifying-otr-deniability/

Proceedings on Privacy Enhancing Technologies 2024(4), 907–928
© 2024 Copyright held by the owner/author(s).
https://doi.org/10.56553/popets-2024-0148

Our new model facilitates new combinations of ideas, e.g. combining malicious adversaries [23] with a big brother distinguisher [10], thereby augmenting the design space for deniability.

Starting a conversation in Signal. Observe that the Signal protocol allows starting a conversation with offline users through the asynchronicity of the initial handshake: Bob prepares a so-called pre-key bundle and uploads it to Signal's key server. To initiate a conversation with Bob, Alice retrieves his pre-key bundle from the key server and sends a message to Bob. (See Figure 3 on page 914 for an algorithmic description of X3DH and PQXDH.)

The X3DH protocol uses the Diffie–Hellman (DH) key exchange, where each user has DH key pairs of different lifetimes (long-term, semi-static, and ephemeral). Bob's pre-key bundle contains his public keys (of all three lifetimes) and a signature certifying his semi-static key under the long-term key. Upon retrieving Bob's pre-key bundle, Alice verifies the signature and generates a fresh ephemeral DH key pair. She computes four (or in case Bob's ephemeral keys run out, only three) DH shared secrets by combining DH keys of different lifetimes, uses the resulting session key to encrypt her first user message with an AEAD scheme, and sends her ephemeral public key and the AEAD ciphertext to Bob. Bob computes the session key in the same manner, decrypts the AEAD ciphertext, and aborts if decryption fails.

The new PQXDH protocol does the same operations as X3DH and adds a KEM encapsulation to achieve post-quantum confidentiality: Bob includes an ephemeral KEM public key in his pre-key bundle and a signature for his KEM key (again under Bob's long-term key). If the ephemeral KEM keys run out, a semi-static KEM key and the corresponding signature are instead included in the pre-key bundle. Alice verifies the signatures on the semi-static DH key and the KEM key and encapsulates against Bob's KEM key. She derives the session key from the DH shared secrets and the KEM shared secret, and encrypts her first user message with the AEAD scheme under the session key. She sends the resulting (KEM and AEAD) ciphertexts and the public key of her freshly generated ephemeral DH key pair to Bob. Bob decapsulates the KEM ciphertext. Bob computes the session key in the same manner, decrypts the AEAD ciphertext, and aborts if decryption fails.

In order to show deniability, we need to specify a Fake algorithm that simulates protocol messages and the session key indistinguishably. To simulate Bob's message, the Fake algorithm needs to produce signatures under Bob's long-term key without Bob's long-term secret key. For signatures on a semi-static key, the Fake algorithm can reuse any signature from the auxiliary info (in practice, this can be a transcript that the attacker learned from eavesdropping or a pre-key bundle) or from an independent session with Bob. For PQXDH, the Fake algorithm needs a signature on an ephemeral key, i.e. a signature that the distinguisher is not aware of. In Section 4 we call these *private signatures*.

To simulate the session key against malicious adversaries, the Fake algorithm requires all DH shared secrets – and for PQXDH additionally the KEM shared secret. In particular, the Fake algorithm does not know the DH secret keys (some are known to the honest party and some to the adversary); following [56], we extract the DH shared secrets from the adversary under a knowledge of DH assumption (see Section 2.2 for details; assume if a party produces a DH public key, this party can produce a DH shared secret with this public key, or nobody can). On Alice's side, the Fake algorithm learns the KEM shared secret by encapsulating; on Bob's side the corresponding KEM secret key is not available for decapsulation. Hence, the Fake algorithm extracts the KEM shared secret from the adversary under the plaintext awareness assumption (see Definition 2.2 for details; assume anybody who produces an encapsulation knows the encapsulated secret).

1.1 Related Work

Built on the literature for deniable authentication [21, 22, 26, 36, 46], Di Raimondo, Gennaro, and Krawczyk [23] have initiated the formal study of deniable key exchange: A distinguisher cannot conclude whether some evidence stems from an actual protocol execution between two parties or whether it was produced by some other means. Many papers have considered deniability for key exchange [6–8, 25, 28, 32–34, 37, 39, 41, 58, 59]. Deniability was also named *repudiability*, mostly in the context of Off-the-Record (OTR) messaging [5, 53–55]. Next, we revisit some more prominent work on deniability notions in general and their application to Signal's initial handshake.

1.1.1 (Non-)Interactive distinguisher. Deniable authentication [21, 26] and some forms of deniable key exchange [23] are defined with respect to a distinguisher[3] that is presented with evidence after the fact, i.e. an *offline distinguisher*. Dodis, Katz, Smith, and Walfish [25] have introduced the notion of an interactive, *online distinguisher*. Unger and Goldberg [54] argue that online deniability is not achievable for an asynchronous key exchange protocol (i.e. one party uploads pre-keys to a central server) with forward secrecy.[4] Hence, in this paper we write *deniability* and refer to offline deniability, unless explicitly stated otherwise.

1.1.2 Deniable Key Exchange. Di Raimondo et al. [23] have introduced *concurrent deniability* and *partial deniability* of key exchange protocols, following the simulation-based approach of deniable authentication. They further share the observation of Pass [46] that a simulator for deniability needs to be implementable; in particular, the simulator needs to refrain from rewinding the adversary, reprogramming a random oracle, and similar techniques for simulation in thought experiments. Their deniability definition was since used in [56] and slightly adapted in [17, 30]. Dagdelen, Fischlin, Gagliardoni, Marson, Mittelbach, and Onete [19] have introduced the first game-based definition for a weak version of deniability of key exchange, named *outsider deniability.* Dodis et al. [25] defined online deniability for key exchange in the Generalized UC framework, which was further refined by Unger and Goldberg [54, 55]. Several works [17, 32–34] consider (adaptive) corruptions in the sense that the attacker, the simulator, and possibly the distinguisher learn the corrupted party's secrets. Appendix A discusses the relation between notions of prior work and our new model in more detail.

1.1.3 Signal's initial handshake. Until recently, Signal employed X3DH [43] as initial handshake protocol, whose security can be

[3]Prior work uses the term *judge*, which we avoid for ambiguity with the legal profession.

[4]They call this assumed impossibility "Iron Triangle".

based on the GapDH assumption [14]. Hashimoto, Katsumata, Kwiatowski, and Prest [30] have proposed a generic post-quantum construction for Signal's initial handshake based on KEMs for confidentiality and either signatures (named SC-AKE), or ring signatures (SC-DAKE), or ring signatures and NIZKs (SC-DAKE') for authentication and increasing levels of deniability. In concurrent work Brendel, Fiedler, Günther, Janson, and Stebila [10] have proposed a similar generic construction named SPQR based on KEMs and designated verifier signatures (DVS). Dobson and Galbraith [24] adopt the X3DH handshake to isogenies, resulting in a protocol named SI-X3DH, which was broken by the SIDH attack [13, 40, 49]. Collins, Huguenin-Dumittan, Nguyen, Rolin, and Vaudenay [16] have proposed K-Waay, which is based on a primitive named split-KEMs that was introduced by [11]. In 2023, Signal deployed PQXDH [38], which uses a more conservative hybrid approach: It combines the classically secure X3DH handshake with a post-quantum secure KEM. This modification aims to add confidentiality against future quantum attackers but is not concerned with quantum attacks against authentication or deniability. Bhargavan, Jacomme, Kiefer, and Schmidt [1, 2] formally verified PQXDH. Fiedler and Günther [27] provide a reductionist key indistinguishability analysis of PQXDH, adding coverage for maximum exposure security to the analysis of [1, 2].

1.1.4 Deniability of Signal's initial handshake. Vatandas, Gennaro, Ithurburn, and Krawczyk [56] have shown that Signal's initial handshake X3DH is deniable in the model of [23] under a novel knowledge of DH assumption. Dobson and Galbraith [24] sketch a similar deniability argument for SI-X3DH. Hashimoto et al. [30] show deniability of SC-DAKE and SC-DAKE' under the model of [23] against semi-honest adversaries and malicious adversaries, respectively, assuming plaintext awareness of the KEM and a related knowledge assumption. Brendel et al. [9, 10] show that SPQR is deniable under a new deniability notion (1-out-of-2 deniability against semi-honest adversaries in the big brother model), which they model after the intuition of Signal's specification [43]. Collins et al. [16] show deniability of K-Waay under a new deniability notion close to the one of Brendel et al. [9, 10].

1.1.5 Deniable Secure Messaging. A protocol combining a key exchange with a subsequent exchange of user messages is called *Secure Messaging*. Unger, Dechang, Bonneau, Fahl, Perl, Goldberg, and Smith [53] cover several forms of deniability for Secure Messaging; among them the distinction between *participation deniability* (a user can deny participation in a specific session) and *message deniability* (a user can deny having written a particular (user) message). Cremers and Zhao [18] extended the game-based deniability model of [10] for key exchange to cover (extended) secure messaging. Reitinger, Malkin, Akgul, Mazurek, and Miers [48], as well as Yadav, Gosain, and Seamons [57] study the social implications of cryptographic deniability for secure messaging.

1.2 Contributions

To capture the precise deniability guarantees of PQXDH, we introduce a new, more expressive deniability model for key exchange protocols (cp. Section 3) that facilitates combinations of previous deniability notions. Our new, game-based deniability model is designed for, but not limited to, asynchronous key exchange. Our model is parameterized to cover deniability notions for a multitude of scenarios: circumstantial knowledge (auxiliary info aux), several adversarial capabilities (malicious, semi-honest, and combinations), how much help the Fake algorithm needs to simulate (partial deniability, or only the peer can simulate, i.e. 1-out-of-2 deniability), and distinguisher capabilities (big brother model, knowledge of auxiliary info). The flexibility of our model allows for easy comparisons of deniability notions and easy adaptability of a proof of deniability to another notion (simply by changing the four aforementioned parameters while keeping the structure of the game, and, hence, the proof).

Furthermore, we provide the first deniability analysis of Signal's new initial handshake PQXDH (cp. Section 4). We show that—due to the signed ephemeral key on Bob's side—the deniability guarantees of PQXDH fall short of those of X3DH. Assuming plaintext awareness of the KEM, which PQXDH adds onto X3DH, and knowledge of a signature hidden from the distinguisher, PQXDH retains the deniability guarantees of X3DH. Table 2 gives the full results.

2 PRELIMINARIES

In this section we review necessary preliminaries. Due to space constraints we defer details on signatures, Diffie–Hellman key exchange, KEMs, and AEAD to Appendix B.

2.1 Notation

We denote the empty list by [], append *element* to *list* by $list \xleftarrow{+} element$, and append the content of $list2$ to $list1$ by $list1 \xleftarrow{+} list2$. In abuse of notation, we allow combining appends, e.g. $(l1, l2) \xleftarrow{+} (l3, l4)$ returns $((l1 \xleftarrow{+} l3), (l2 \xleftarrow{+} l4))$, and allow the same syntax for adding to sets. We denote by $a, b, c \in S$ that S contains a, b, c. Given a map a, we refer to its entry at position i with $a[i]$.

Two distributions are said to be $(t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$-indistinguishable if no algorithm running in time $t_{\mathcal{D}}$ has an advantage better than $\epsilon_{\mathcal{D}}$ in distinguishing the two distributions. We use $y \leftarrow\$ A(x)$ to denote the random output y of algorithm A for input x, where the probability is over A's internal randomness. We use the arrow $\leftarrow$ for any assignment statements.

2.2 Assumption on Diffie–Hellman Key Exchange

We express whether a party can compute functions of its own Diffie–Hellman (DH) secret key with the following assumption: Consider an algorithm $\mathcal{A}$ that outputs a DH public key. Then an extractor $\mathcal{E}_{\mathcal{A}}^{DH}$ with runtime $t_{\mathcal{E}}$ given the same inputs as $\mathcal{A}$ and an extra DH public key can extract the DH shared secret of both public keys from $\mathcal{A}$. If $\mathcal{E}_{\mathcal{A}}^{DH}$ fails to extract the shared DH secret, then no other algorithm with runtime $t_{\mathcal{D}}$ succeeds with probability better than $\epsilon_{\mathcal{D}}$. To put it in a nutshell, extraction succeeds within $t_{\mathcal{E}}$, or not even within $t_{\mathcal{D}}$. We denote the DH shared secret between two DH public keys U, Z with respect to some generator g as $DH(U, Z)$.

Observe that in [56], Vatandas et al. show that the knowledge of exponent assumption and their novel knowledge of discrete logarithm assumption imply the knowledge of DH (KDH) assumption.

```
$\mathcal{G}^{dec}_{\mathsf{KEM}}(C, \mathcal{D})$:
for $i \in [n_k]$
    $(\mathsf{pk}_i, \mathsf{sk}_i) \leftarrow\$ \mathsf{KEM.KGen}()$
$\vec{\mathsf{pk}} \leftarrow \{\mathsf{pk}_i\}_{i \in [n_k]}$
$r \leftarrow\$ \mathcal{R}_C$
$v \leftarrow C^{\textsc{Decaps}}(\vec{\mathsf{pk}}; r)$
return $\mathcal{D}(v)$

$\textsc{Decaps}(i, ct)$:
return $\mathsf{KEM.Dec}(\mathsf{sk}_i, ct)$

$\mathcal{G}^{ext}_{\mathsf{KEM}}(C, \mathcal{E}^{PA}_C, \mathcal{D})$:
for $i \in [n_k]$
    $(\mathsf{pk}_i, \mathsf{sk}_i) \leftarrow\$ \mathsf{KEM.KGen}()$
$\vec{\mathsf{pk}} \leftarrow \{\mathsf{pk}_i\}_{i \in [n_k]}$
$r \leftarrow\$ \mathcal{R}_C$
$v \leftarrow C^{\textsc{Decaps}}(\vec{\mathsf{pk}}; r)$
return $\mathcal{D}(v)$

$\textsc{Decaps}(i, ct)$:
return $\mathcal{E}^{PA}_C(i, ct, r)$
```

Figure 1: Games for plaintext awareness, see Definition 2.2.

Compared to the EKDH assumption, the algorithm $\mathcal{A}$ of the KDH assumption receives the public key U as input.

Definition 2.1 (EKDH Assumption [56]). Let G be a cyclic group with generator g and AuxPrep a sampler for auxiliary inputs. Let $\mathcal{A}$ be any algorithm running in time $t_{\mathcal{A}}$ which runs on input aux where aux $\leftarrow\$$ AuxPrep, and outputs $Z \in G$; we denote with $Z = \mathcal{A}(\mathsf{aux}, r)$ the output of running $\mathcal{A}$ on input aux with coins r.

We say that the $(t_{\mathcal{A}}, t_{\mathcal{E}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$-Extended Knowledge of DH (EKDH) Assumption holds over group G and for auxiliary info AuxPrep, if for every such $\mathcal{A}$, there exists a companion algorithm $\mathcal{E}^{DH}_{\mathcal{A}}$ (called the extractor for $\mathcal{A}$) such that: $\mathcal{E}^{DH}_{\mathcal{A}}$ runs on input aux, r and an additional input $U \in G$ in time $t_{\mathcal{E}}$ and outputs $\hat{Z} \in G$ or $\perp$ such that

- If $\mathcal{E}^{DH}_{\mathcal{A}}(U, \mathsf{aux}, r) \neq \perp$ then $\hat{Z} = DH(U, Z)$
- For every algorithm C running in time $t_{\mathcal{D}}$, we have that $\Pr[C(U, r, \mathsf{aux}) = DH(U, Z) \mid \mathcal{E}^{DH}_{\mathcal{A}}(U, \mathsf{aux}, r) = \perp] \leq \epsilon_{\mathcal{D}}$, where $Z = \mathcal{A}(\mathsf{aux}, r)$ and the probability is taken over the coins of C and uniform distribution on r.

2.3 Key Encapsulation Mechanisms (KEMs)

In Appendix B.2, we provide a formal definition of KEMs. We define the key-collision probability γ_{coll} for KEM as the probability of the public keys of two independently sampled key pairs to coincide.

Plaintext awareness for KEMs denotes whether one can create a valid KEM encapsulation without knowing the corresponding shared secret. Based on plaintext awareness for PKE [3, 4], and for KEMs [20, 35], Hashimoto et al. [30] have adapted the definition to plaintext awareness for KEMs in the multi-key setting. We adapt their definition from asymptotic security to concrete security.

Definition 2.2 (Plaintext Awareness for KEMs [30]). A KEM scheme KEM is $(n_k, t_C, t_{\mathcal{E}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$-plaintext aware (PA1-secure) if for all (non-uniform) ciphertext creators C running in time t_C, there exists an efficient extractor $\mathcal{E}^{PA}_C$ running in time $t_{\mathcal{E}}$ such that for any distinguisher $\mathcal{D}$ running in time $t_{\mathcal{D}}$, the two experiments $\mathcal{G}^{dec}_{\mathsf{KEM}}(C, \mathcal{D})$ and $\mathcal{G}^{ext}_{\mathsf{KEM}}(C, \mathcal{E}^{PA}_C, \mathcal{D})$ in Figure 1 are indistinguishable except with a probability of $\epsilon_{\mathcal{D}}$.

Intuitively, a KEM is plaintext aware if creating a KEM ciphertext implies knowledge of the corresponding shared secret. This is formally modeled with a ciphertext creator who has access to a decryption oracle. This decryption oracle is implemented either with the actual decryption or with an extractor that depends on the ciphertext creator and its randomness. Finally, the ciphertext creator outputs a string, which is fed into a distinguisher. The distinguisher needs to distinguish how the decryption oracle is implemented.

Note that Kyber [51] using the FO transform with explicit rejection instead of implicit rejection is plaintext aware.

2.4 Key Exchange Protocols

We review the syntax for key exchange protocols, essentially following [9, 10]. We differentiate between protocol messages m and user (plaintext) messages μ (such as "Hi Bob, how are you doing?"). A key exchange protocol may allow a user message to be sent along with a protocol message during the execution, e.g. X3DH and PQXDH.

Definition 2.3 (Key Exchange Protocol). A 2-party key exchange protocol is a triple of algorithms KE = (KGenLT, KGenSS, Run):

- $\mathsf{KGenLT}() \xrightarrow{\$} (ltpk_U, ltsk_U)$: A probabilistic long-term key generation algorithm that outputs a public-key/secret-key pair attributed to user U.
- $\mathsf{KGenSS}() \xrightarrow{\$} (sspk^{\mathsf{ssid}}_U, sssk^{\mathsf{ssid}}_U)$: A probabilistic semi-static key generation algorithm that outputs a public-key/secret-key pair attributed to user U and semi-static identifier ssid.
- $\mathsf{Run}(ltsk_U, \vec{sssk}_U, \vec{ltpk}, \vec{sspk}, \pi, m, \mu) \xrightarrow{\$} (\pi', m', \mu')$: A probabilistic session execution algorithm that takes as input a party's long-term secret key $ltsk_U$, a list of that party's semi-static secret keys $\vec{sssk}_U$, lists of long-term and semi-static public keys for all honest parties $\vec{ltpk}$ and $\vec{sspk}$, a session state π, an incoming message m, and a (potentially empty) user message μ to be sent, and outputs an updated session state π', a (possibly empty) outgoing message m', and a (possibly empty) incoming user message μ'.

We assume a distinguished message $m = (\mathsf{create}, \mathsf{info}_{\mathsf{create}})$ upon which Run sets up the session and produces the first message. Note that an ephemeral key generation algorithm is not mandatory and can happen inside of Run.

Parties and sessions. Let $\mathcal{P}$ be the set of n_p parties, each of whom has a long-term public-key/secret-key pair generated by an algorithm KGenLT. Each party may run multiple instances of the protocol simultaneously or sequentially, each of which is called a session. The ith session at party P is denoted π^i_P. For each session, the party maintains the following collection of session-specific information:

- $\mathsf{oid} \in \mathcal{P}$: The identity of the session owner.
- $\mathsf{pid} \in \mathcal{P} \cup \{\star\}$: The identity of the intended peer, which may initially be unknown (indicated by $\star$).
- role $\in$ {initiator, responder}: The role of the party in this session.
- $\mathsf{K} \in \mathcal{K}_{\mathsf{KE}} \cup \{\perp\}$: The session key established in this session, initialized to $\perp$.
- $\mathsf{info}_{\mathsf{create}}$ is a tuple of elements that indicates the components of the first message. We use the following elements:
  - $\mathsf{ssid} \in [n_{ss}]$ denotes the identifier of the initiator's semi-static key in this session. If π.role = initiator this refers to $sspk^{\mathsf{ssid}}_{\pi.\mathsf{oid}}$, if π.role = responder this refers to $sspk^{\mathsf{ssid}}_{\pi.\mathsf{pid}}$.
  - $e_{\mathsf{DH}} \in \{\mathsf{true}, \mathsf{false}\}$ denotes whether an ephemeral DH key is included in the first message.
  - $e_{\mathsf{KEM}} \in \{\mathsf{true}, \mathsf{false}\}$ denotes whether an ephemeral KEM key is included in the first message.

Asynchronous key exchange. In principle, a key exchange protocol can have an arbitrary number of message flows n_m, which correspond to multiple calls to Run for a single session. In normal execution of an *asynchronous* authenticated key exchange protocol, the following three calls to Run occur: 1) a call to Run at the initiator Bob with $m = (\mathsf{create}, \mathsf{info}_{\mathsf{create}})$, which sets up the initiator session and outputs the pre-key bundle, possibly including ephemeral public keys; 2) a call to Run at the responder Alice with the initiator's pre-key bundle, which generates a session key and outputs a key exchange message; and 3) a call to Run at the initiator with the responder's long-term public key and key exchange message, which generates a session key and has no output message. If a party outputs an empty message before the protocol execution is finished, it aborts the protocol.

3 OUR DENIABILITY MODEL

We introduce our deniability model that facilitates combining ideas of deniability definitions from prior literature, hence gaining expressiveness. We show how our model can match definitions from prior literature in Appendix A.

We want to capture that artifacts of a key exchange between Alice and Bob do not convince a third party that the key exchange actually took place. Consider Alice, who tries to prove Bob's involvement. To defend against Alice's claim, Bob must be able to argue that the artifacts were produced without him, e.g. by Alice or by just anybody. Hence, we require the existence of a Fake algorithm that produces these artifacts, where the artifacts are indistinguishable from artifacts of an actual execution of the key exchange protocol.

Our model captures this with the following game: First, the challenger prepares the game by initializing variables, sampling keys for all users, preparing auxiliary info aux according to AuxPrep, and sampling a secret bit b. Second, the adversary interacts with some oracles. One of the oracles depends on the secret bit b and embeds the challenge: To distinguish whether (a part of) a protocol transcript was generated honestly or with a dedicated Fake algorithm, which needs to be given in the proof. Third, the distinguisher guesses the challenge bit, based on the previous actions of the adversary. In particular, the distinguisher gets the transcript of the interactions between the adversary and the challenge oracle and the (honestly computed) session keys for all sessions. Note that $\mathcal{A}$ and Fake get access to the auxiliary info aux.

The game is parameterized by the adversary's capabilities O_A, i.e. the oracles it can query, the capabilities of the Fake algorithm O_F, the capabilities of the distinguisher O_D, and the sampling method for auxiliary information AuxPrep. Extending O_A or O_D strengthens the deniability guarantee, while extending O_F or AuxPrep weakens the guarantee. In Sections 3.1 to 3.4 we explain these parameters in detail. In Section 3.5 we discuss how to compare definitions with different parameters. The flexibility and expressiveness of these parameters is the novelty of our model.

Definition 3.1 (Deniability for Key Exchange). We say that a key exchange protocol $\mathsf{KE} = (\mathsf{KGenLT}, \mathsf{KGenSS}, \mathsf{Run})$ is $(n_p, n_{ss}, q_{\mathrm{O}_A}, q_{\mathrm{O}_F}, q_{\mathrm{O}_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$-deniable wrt. $(\mathrm{O}_A, \mathrm{O}_F, \mathrm{O}_D)$-oracles, where $\mathrm{O}_A \subseteq \{\textsc{Reg}, \textsc{Init}, \textsc{ChallInit}, \textsc{ChallResp}, \textsc{RegHon}, \textsc{ChallHonInit}, \textsc{ChallHonResp}\}$, $\mathrm{O}_F \subseteq \{\mathsf{SK}, \textsc{User}_{x,y}\}$, $\mathrm{O}_D \subseteq \{\mathsf{SKs}, \textsc{aux}\}$, and auxiliary inputs sampled with AuxPrep, if for any adversary $\mathcal{A}$ running in time $t_{\mathcal{A}}$, making queries to O_A limited by q_{O_A}, there exists an algorithm Fake making queries to O_F limited by q_{O_F} such that for any distinguisher $\mathcal{D}$ making queries to O_D limited by q_{O_D} and running in time $t_{\mathcal{D}}$, it holds that $\Pr[\mathcal{G}^{\mathrm{O}_A,\mathrm{O}_F,\mathrm{O}_D\text{-}den}_{\mathsf{KE},\mathsf{AuxPrep},n_p,n_{ss}}(\mathcal{A},\mathcal{D}) = 1] \leq \frac{1}{2} + \epsilon_{\mathcal{D}}$, where $\mathcal{G}^{\mathrm{O}_A,\mathrm{O}_F,\mathrm{O}_D\text{-}den}_{\mathsf{KE},\mathsf{AuxPrep},n_p,n_{ss}}(\mathcal{A},\mathcal{D})$ is defined in Figure 2.

Type of Deniability	O_A	O_F	O_D
against malicious adv.	$\textsc{Reg}, \textsc{Init}, \textsc{Chall}$	$\Diamond$	$\Diamond$
against semi-honest adv.	$\textsc{RegHon}$, $\textsc{ChallHon}$	$\Diamond$	$\Diamond$
against passive adv.	$\textsc{ChallHon}$	$\Diamond$	$\Diamond$
partial deniability wrt. an x, y oracle	$\Diamond$	$\textsc{User}_{x,y}$	$\Diamond$
1-out-of-2 deniability	$\Diamond$	SK	$\Diamond$
in the big brother model	$\Diamond$	$\Diamond$	SKs
with aux known to the distinguisher	$\Diamond$	$\Diamond$	AUX

Table 1: Translating oracles into names for adversarial capabilities and deniability guarantees. We refer to an arbitrary set with $\Diamond$. For the challenge oracles we omit the suffix indicating which role they act as.

Intuitively, the Fake algorithm models that a protocol message was not necessarily produced by the alleged sender. Consider an adversary $\mathcal{A}$ acting as Alice (i.e. the adversary generates keys for Alice), who produces Alice's message(s) according to Run and Bob's message(s) and the session key according to the Fake algorithm. Note that the Fake algorithm also needs to produce the session key to ensure deniability of the subsequent protocol. Since the Fake algorithm is public, Bob can argue that the transcript and session key were produced by Alice using the Fake algorithm, *without* his involvement. On a technical note, the Fake algorithm knows the code and randomness of the adversary $\mathcal{A}$ since they are in cahoots. This allows the Fake algorithm to extract knowledge from the adversary under some knowledge assumptions.

In the following we describe the choices for using our model and how our model allows easy comparisons of deniability notions.

3.1 AuxPrep: Sampling auxiliary information

The auxiliary info aux models that an adversary may obtain relevant information prior to the experiment. To the best of our knowledge, this idea was introduced by [23]: They use aux to model that the adversary may have obtained transcripts of previous protocol executions by eavesdropping. Following [23], the auxiliary info is also available to the Fake algorithm, i.e. anybody who knows the auxiliary info can simulate a transcript and session key. Optionally, the distinguisher gets access to aux as well (cp. Section 3.4.2).

3.2 O_A: capabilities of $\mathcal{A}$

The oracles in O_A model the behavior of the adversary, i.e. whether it is malicious or semi-honest,[5] and define the challenge. In the following, we remark on how we handle the roles of parties and detail the usage of the respective subsets of oracles.

[5]In Appendix C we also treat passive adversaries for the sake of completeness. We do not consider passive adversaries for our analysis and therefore defer it to the appendix.

```
$\mathcal{G}^{O_A,O_F,O_D\text{-}den}_{\mathsf{KE},\mathsf{AuxPrep},n_p,n_{ss}}(\mathcal{A},\mathcal{D})$:
1  $Q[\cdot] \leftarrow []$;   $K[\cdot] \leftarrow \perp$;   $C \leftarrow \emptyset$
2  $(\vec{\mathsf{pk}}, \vec{\mathsf{sk}}) \leftarrow\$ \mathsf{KeyPrep}(n_p, n_{ss})$
3  $\mathsf{aux} \leftarrow\$ \mathsf{AuxPrep}(\vec{\mathsf{pk}}, \vec{\mathsf{sk}})$
4  $b \leftarrow\$ \{0,1\}$
5  $r \leftarrow\$ \mathcal{R}_{\mathcal{A}}$ //randomness for the adversary
6  $\mathcal{A}^{O_A}(\vec{\mathsf{pk}}, \mathsf{aux}; r)$
7  $b' \leftarrow\$ \mathcal{D}^{O_D}(\vec{\mathsf{pk}}, r, Q, K)$
8  return $[\![ b' = b ]\!]$

$\mathsf{KeyPrep}(n_p, n_{ss})$:
9  for $U \in [n_p]$ //number of keys to prepare
10     $(ltpk_U, ltsk_U) \leftarrow\$ \mathsf{KGenLT}()$
11     for $\mathsf{ssid} \in [n_{ss}]$
12        $(sspk_U^{\mathsf{ssid}}, sssk_U^{\mathsf{ssid}}) \leftarrow\$ \mathsf{KGenSS}()$
13 $\vec{ltpk} \leftarrow \{ltpk_U\}_{U \in [n_p]}$;   $\vec{sspk} \leftarrow \{sspk_U^{\mathsf{ssid}}\}_{U \in [n_p]}^{\mathsf{ssid} \in [n_{ss}]}$
14 $\vec{ltsk} \leftarrow \{ltsk_U\}_{U \in [n_p]}$;   $\vec{sssk} \leftarrow \{sssk_U^{\mathsf{ssid}}\}_{U \in [n_p]}^{\mathsf{ssid} \in [n_{ss}]}$
15 return $((\vec{ltpk}, \vec{sspk}), (\vec{ltsk}, \vec{sssk}))$

$\textsc{Reg}(U, \vec{\mathsf{pk}}_U)$:
16 if $U \in [n_p] \cup C$ //abort if another key is already known for $U$
17    return $\perp$
18 $C \leftarrow C \cup \{U\}$
19 $\vec{\mathsf{pk}} \xleftarrow{+} \vec{\mathsf{pk}}_U$ //includes semi-static keys
20 return success

$\textsc{Init}(U, s, \mathsf{role}, V)$:
21 if $\pi_U^s = \perp \wedge U \notin C$
22    $\pi_U^s.\mathsf{role} \leftarrow \mathsf{role}$
23    $\pi_U^s.\mathsf{oid} \leftarrow U$ //set owner identity
24    $\pi_U^s.\mathsf{pid} \leftarrow V$ //set partner identity (★ if post-specified peers)
25    return success
26 return $\perp$

$\textsc{RegHon}(U)$:
27 if $U \in [n_p] \cup C$ //abort if another key is already known for $U$
28    return $\perp$
29 $(\vec{\mathsf{pk}}_U, \vec{\mathsf{sk}}_U) \leftarrow\$ \mathsf{KeyPrep}(1, n_{ss})$
30 $\vec{\mathsf{pk}} \xleftarrow{+} \vec{\mathsf{pk}}_U$;   $\vec{\mathsf{sk}} \xleftarrow{+} \vec{\mathsf{sk}}_U$
31 $C \leftarrow C \cup \{U\}$
32 return $(\vec{\mathsf{pk}}_U, \vec{\mathsf{sk}}_U)$ //simulates $\mathcal{A}$ honestly generating keys

$\textsc{ChallHonInit}(U, V, \mathsf{info}_{\mathsf{create}}, \vec{\mu}, r_C)$:
33 if $U \in C$ //do not allow challenging a party under $\mathcal{A}$'s control
34    return $\perp$
35 $\pi_U.\mathsf{oid} \leftarrow U$;   $\pi_U.\mathsf{pid} \leftarrow V$
36 $\pi_V.\mathsf{oid} \leftarrow V$;   $\pi_V.\mathsf{pid} \leftarrow U$
37 $\pi_U.\mathsf{role} \leftarrow \mathsf{initiator}$;   $\pi_V.\mathsf{role} \leftarrow \mathsf{responder}$
38 $m_1 \leftarrow (\mathsf{create}, \mathsf{info}_{\mathsf{create}})$
39 $(\mu_1, \ldots, \mu_{n_m}) \leftarrow \vec{\mu}$
40 for $i \in [1, 3, \ldots, n_m - 1]$ //until $n_m$ if $n_m$ is odd
41    if $b = 0$
42       $(\pi_U, m_{i+1}) \leftarrow\$ \mathsf{Run}(\mathsf{sk}_U, \vec{\mathsf{pk}}, \pi_U, m_i, \mu_i)$
43    else
44       $(\pi_U, m_{i+1}) \leftarrow\$ \mathsf{Fake}^{O_F}(\vec{\mathsf{pk}}, \pi_U, m_i, \mu_i, \mathsf{aux}, r_C)$
45    $(\pi_V, m_{i+2}) \leftarrow \mathsf{Run}(\mathsf{sk}_V, \vec{\mathsf{pk}}, \pi_V, m_{i+1}, \mu_{i+1}; r_C)$
46 $Q[\pi_U] \leftarrow (m_i)_{i=1}^{n_m}$;   $K[\pi_U] \leftarrow \pi_U.K$
47 return $(Q[\pi_U], K[\pi_U])$

$\textsc{ChallHonResp}(U, V, \mathsf{info}_{\mathsf{create}}, \vec{\mu}, r_C)$:
48 if $V \in C$ //do not allow challenging a party under $\mathcal{A}$'s control
49    return $\perp$
50 $\pi_U.\mathsf{oid} \leftarrow U$;   $\pi_U.\mathsf{pid} \leftarrow V$
51 $\pi_V.\mathsf{oid} \leftarrow V$;   $\pi_V.\mathsf{pid} \leftarrow U$
52 $\pi_U.\mathsf{role} \leftarrow \mathsf{initiator}$;   $\pi_V.\mathsf{role} \leftarrow \mathsf{responder}$
53 $m_1 \leftarrow (\mathsf{create}, \mathsf{info}_{\mathsf{create}})$
54 $(\mu_1, \ldots, \mu_{n_m}) \leftarrow \vec{\mu}$
55 for $i \in [1, 3, \ldots, n_m - 1]$ //until $n_m$ if $n_m$ is odd
56    $(\pi_U, m_{i+1}) \leftarrow \mathsf{Run}(\mathsf{sk}_U, \vec{\mathsf{pk}}, \pi_U, m_i, \mu_i; r_C)$
57    if $b = 0$
58       $(\pi_V, m_{i+2}) \leftarrow\$ \mathsf{Run}(\mathsf{sk}_V, \vec{\mathsf{pk}}, \pi_V, m_{i+1}, \mu_{i+1})$
59    else
60       $(\pi_V, m_{i+2}) \leftarrow\$ \mathsf{Fake}^{O_F}(\vec{\mathsf{pk}}, \pi_V, m_{i+1}, \mu_{i+1}, \mathsf{aux}, r_C)$
61 $Q[\pi_V] \leftarrow (m_i)_{i=1}^{n_m}$;   $K[\pi_V] \leftarrow \pi_V.K$
62 return $(Q[\pi_V], K[\pi_V])$

$\textsc{ChallInit}(U, s, m, \mu)$ $\boxed{\textsc{ChallResp}(U, s, m, \mu)}$:
63 if $\pi_U^s = \perp$ //initialized session only
     $\vee\ \pi_U^s.\mathsf{role} = \mathsf{responder}$ $\boxed{\mathsf{initiator}}$
64    return $\perp$
65 if $b = 0$
66    $(\pi_U^s, m') \leftarrow\$ \mathsf{Run}(\mathsf{sk}_U, \vec{\mathsf{pk}}, \pi_U^s, m, \mu)$
67 else
68    $(\pi_U^s, m') \leftarrow\$ \mathsf{Fake}^{O_F}(\vec{\mathsf{pk}}, \pi_U^s, m, \mu, \mathsf{aux}, r)$
69 $Q[\pi_U^s] \xleftarrow{+} (m, m')$
70 $K[\pi_U^s] \leftarrow \pi_U^s.\mathsf{K}$
71 return $m'$

$\textsc{User}_{x,y}(\pi, u, m, \mu)$:
72 $U \leftarrow \pi.\mathsf{oid}$
73 $H \leftarrow H_\pi^u$ //easier notation for $u$th honest dummy for session $\pi$
74 if $\pi_U^H = \perp$
75    $(\mathsf{pk}_H, \mathsf{sk}_H) \leftarrow\$ \mathsf{KeyPrep}(1, 1)$
76    $\vec{\mathsf{pk}} \leftarrow \vec{\mathsf{pk}} \cup \{\mathsf{pk}_H\}$
77    if $\pi.\mathsf{role} = \mathsf{initiator}$
78       $\pi_U^H.\mathsf{role} \leftarrow x$
79    else
80       $\pi_U^H.\mathsf{role} \leftarrow y$
81    $\pi_U^H.\mathsf{oid} \leftarrow U$
82    $\pi_U^H.\mathsf{pid} \leftarrow H$
83 $(\pi_U^H, m', \mu') \leftarrow\$ \mathsf{Run}(\mathsf{sk}_U, \vec{\mathsf{pk}}, \pi_U^H, m, \mu)$
84 return $m'$

$\mathsf{SK}(\pi)$:
85 return $\mathsf{sk}_{\pi.\mathsf{pid}}$

$\mathsf{SK}\textsc{s}()$:
86 return $\vec{\mathsf{sk}}$

$\textsc{aux}()$:
87 return aux
```

Figure 2: Our deniability game parameterized in the adversary's capabilities $O_A \subseteq \{\textsc{Reg}, \textsc{Init}, \textsc{ChallInit}, \textsc{ChallResp}, \textsc{RegHon}, \textsc{ChallHonInit}, \textsc{ChallHonResp}\}$, the capabilities of the Fake algorithm $O_F \subseteq \{\mathsf{SK}, \textsc{User}_{x,y}\}$, the capabilities of the distinguisher $\mathcal{D}$ $O_D \subseteq \{\mathsf{SK}\textsc{s}, \textsc{aux}\}$, and the sampling algorithm for auxiliary info AuxPrep.

3.2.1 Initiator and receiver deniability. We obtain *initiator deniability* by having $\textsc{ChallInit} \in O_A$ or $\textsc{ChallHonInit} \in O_A$, and *responder deniability* by having $\textsc{ChallResp} \in O_A$ or $\textsc{ChallHonResp} \in O_A$; i.e. we give separate challenge oracles per role. This allows separate analyses of deniability per role. (Section 4 shows that the deniability guarantees for both roles differ for X3DH and PQXDH.) Also, this enables more precise specifications, e.g. deniability for the responder may have stricter requirements than for the initiator.

3.2.2 Malicious adversaries. We obtain *deniability against malicious adversaries* with $O_A = \{\textsc{Reg}, \textsc{Init}, \textsc{ChallInit}, \textsc{ChallResp}\}$, allowing at most q_R queries to the Reg oracle, q_I queries to the Init oracle, q_{CI} queries to the ChallInit oracle, and q_{CR} queries to the ChallResp oracle, for $q_R, q_I, q_{CI}, q_{CR} \in q_{O_A}$. The Reg oracle allows the adversary to register maliciously generated keys with the challenger. The Init oracle allows the adversary to initialize sessions for honest users and to partner these sessions arbitrarily. The ChallInit and ChallResp oracles allow the adversary to send arbitrary messages to sessions (including, e.g. messages replayed from other sessions) and to interleave sessions at will. The ChallInit and ChallResp oracles respond either according to the protocol specification via Run or "simulate" via Fake, depending on the challenge bit b. For the case of Signal's initial handshake, note that this also captures an adversarially controlled server, i.e. a server who forwards messages wrongly. Only sessions of honestly generated users can be challenged since the Init oracle does not create sessions for users with maliciously generated keys.

3.2.3 Semi-honest Adversaries. Next, we obtain the notion of *deniability against semi-honest adversaries* by setting the oracle $O_A = \{\textsc{RegHon}, \textsc{ChallHonInit}, \textsc{ChallHonResp}\}$, allowing at most q_{RH} queries to the RegHon oracle, q_{CI} queries to the ChallHonInit oracle, and q_{CR} queries to the ChallHonResp oracle, for $q_{RH}, q_{CI}, q_{CR} \in q_{O_A}$. To register a key pair for a new user, the adversary can query the RegHon oracle, which returns the user's private key(s) to the adversary. A semi-honest adversary partners sessions correctly and obeys the protocol flow. Hence, we merge the Init oracle and the challenge oracles. These ChallHonInit, ChallHonResp oracles partner the sessions correctly and directly produce the complete transcript, thereby enforcing correct message flow.

Only honest users can be challenged (cp. lines 33 and 48). Note that the adversary provides the randomness to the challenge oracles to create messages "in place of the adversary"[6]. The adversary also provides the protocol-specific $\mathsf{info}_{\mathsf{create}}$ to indicate the components of the first message.

3.2.4 Combining different adversarial capabilities. It is possible to combine the aforementioned oracles , e.g. achieving responder deniability against malicious adversaries and initiator deniability against semi-honest adversaries with malicious keys via $\mathrm{O}_A = \{\textsc{Reg}, \textsc{Init}, \textsc{ChallResp}, \textsc{ChallHonInit}\}$. Note that O_A should always contain at least one challenge oracle; and for the ChallInit, ChallResp oracles the Init oracle is necessary. Oracles for both malicious and semi-honest adversaries seem redundant but may have use cases.

3.2.5 Adaptive Corruptions. Several works [17, 32–34] allow the adversary to adaptively corrupt parties. The attacker, the simulator, and possibly the distinguisher learn the corrupted parties' secrets. Fundamentally, deniability is a statement that a Fake algorithm can simulate a transcript (and session key). The Fake algorithm needs to work even if the adversary does not corrupt any party. Hence, the strategy of the Fake algorithm should not rely on secret keys of corrupted parties and that is why our model does not consider adaptive corruptions. Appendices A.2 and A.8 describe how to represent non-adaptive corruptions in our model.

3.3 O_F: capabilities of Fake

The oracles in O_F are accessed by the Fake algorithm. They offer help in simulating a transcript and session key and should reflect the capabilities of an adversary who wants to frame a user for a particular transcript. If fewer oracles are required, then more people can simulate, i.e. the deniability guarantee gets stronger.

3.3.1 Partial deniability ($\textsc{User}_{x,y}$ oracle (parameterized by x, y)). We obtain *partial deniability wrt. x, y oracles* by having $\textsc{User}_{x,y} \in \mathrm{O}_F$, allowing at most q_U sessions with the $\textsc{User}_{x,y}$ oracle per session π, where $x, y \in \{\text{initiator}, \text{responder}, \perp\}$. Each invocation of the Fake algorithm (lines 44, 60, and 68) has access to exactly one session π. The Fake algorithm can query the $\textsc{User}_{x,y}$ oracle with this session π to get access to another session with $\pi.\mathsf{oid}$ (the owner of π) with the role x (if $\pi.\mathsf{oid}$ is the initiator in π) or y (if $\pi.\mathsf{oid}$ is the responder in π), where $\perp$ denotes no access. The Fake algorithm can have up to $q_U \in q_{\mathrm{O}_F}$ sessions with the $\textsc{User}_{x,y}$ oracle per session π. For $u \in [q_U]$, the user $\pi.\mathsf{oid}$ is partnered with a (freshly generated) honest user H_π^u (and in particular not with $\pi.\mathsf{pid}$), where $H_\pi^u \in \mathcal{H}$ and $\mathcal{H} \cap [n_p] = \emptyset$.

This idea is sourced from *partial deniability* [23] (informally known as *peer independence* [23] or *receiver obliviousness* [30], following the idea of *post-specified peers* [12]).

This oracle models that an adversary can start a separate session with the victim $\pi.\mathsf{oid}$ to produce the transcript. This oracle can be helpful if messages do not contain session-specific data.

3.3.2 1-out-of-2 or 1-out-of-∞ deniability (SK oracle). We obtain *1-out-of-2 deniability* by having $\mathrm{SK} \in \mathrm{O}_F$. If $\mathrm{SK} \notin \mathrm{O}_F$, then we obtain *1-out-of-∞ deniability.* The SK oracle grants the Fake algorithm access to the peer's secret keys, thereby modeling that the peer (and only the peer) can execute the Fake algorithm. Hence, 1-out-of-∞ deniability is more desirable. Note that this refers to long-term and semi-static keys, i.e. keys that are registered with the challenger. In particular, the SK oracle cannot yield secret keys of users that the adversary registered via the Reg oracle.

If a transcript (and session key) can stem from either an honest execution or the peer creating the transcript on his or her own (with the Fake algorithm) then there are two options for this; hence, the name 1-out-of-2 deniability. Absence of this oracle indicates that anybody could have produced a transcript, hence, we dub it 1-out-of-∞ deniability. These terms were coined by [29] for authentication protocols[7] and adopted for key exchange by [10].

The idea of using one party's secret key for faking a transcript was also used by [56], though coming from a different angle: They allow the simulator (roughly corresponding to our Fake algorithm) to use the adversary's long-term secret key to fake a transcript. They justify this with the adversary having registered its long-term key to some form of PKI, possibly with an extractable proof of knowledge, from which the simulator can learn the secret key.

3.4 O_D: capabilities of $\mathcal{D}$

The distinguisher always gets all users' public keys, the transcripts and session keys from the interaction with the Chall oracle, and the randomness of the adversary. The oracles in O_D define the extra power that the distinguisher may have in the form of access to all (honestly generated) secret keys (i.e. the big brother model) or the auxiliary info. The more oracles the distinguisher has access to, the stronger the deniability guarantee we obtain.

3.4.1 Big brother model. We obtain *deniability in the big brother model* by having $\mathrm{SKs} \in \mathrm{O}_D$. By querying the SKs oracle, the distinguisher gets access to all secret keys that the challenger prepared for all users as well as secret keys registered by the adversary via the RegHon oracle. This models that the distinguisher can subpoena all parties into giving up their private keys or can exert social pressure in a different manner. The SKs oracle does not yield secrets of keys registered via the Reg oracle—the challenger does not know these secret keys and possibly the adversary does not either. Similarly, the SKs oracle does not yield ephemeral secret keys, as they are not registered with the challenger (and, for X3DH and PQXDH, honest parties erase ephemeral secret keys directly after use).

The term *big brother* was introduced by [44] in the context of deniable ring authentication. The adversary divulging its secret key to the distinguisher was coined complete deniability by [41]. Informal descriptions [53, 54] and the UC-based definitions of [52] are set in this model without using this name. The game-based definition of [10] first uses the term big brother model for deniability of key exchange protocols.

3.4.2 Auxiliary info known to the distinguisher. We obtain *deniability with* aux *known to the distinguisher* by having $\textsc{aux} \in \mathrm{O}_D$. By querying the aux oracle, the distinguisher gets access to the auxiliary information (cp. Section 3.1). This models that the distinguisher learns whatever the adversary has eavesdropped on before the beginning of the experiment. If $\textsc{aux} \notin \mathrm{O}_D$ then the deniability guarantee *relies on* the distinguisher never learning aux. This is

[6]In consequence, the distinguisher learns this randomness as well.

[7]They also call a ring signature with a ring of n members 1-out-of-n deniable. This allows for a philosophical debate whether this applies to messaging services with n registered users. We argue that anybody can sign up for the service and, hence, there is no practical difference to 1-out-of-∞ deniability.

important for analyzing PQXDH, where the Fake algorithm reuses signatures from aux, which are not reused in honest executions, and thereby allows the distinguisher to tell Run and Fake apart.

3.5 Comparability of deniability notions

Our model allows fairly easy comparisons between any two deniability notions by weighing $\mathrm{O}_A, \mathrm{O}_F, \mathrm{O}_D$, AuxPrep between the two notions. If the adversary $\mathcal{A}$ or the distinguisher $\mathcal{D}$ gain more capabilities, i.e. if $\mathrm{O}_A, \mathrm{O}_D$ contain more or stronger oracles, the deniability guarantee becomes stronger (since it holds in more cases). Though, the AUX oracle does not make a difference in case aux $= \perp$. However, if Fake gains more capabilities, i.e. if O_F contains more oracles, the deniability guarantee becomes weaker: Simulating a transcript requires access to the oracles in O_F. Similarly, the more information is included in the auxiliary information aux, the harder it is to match these circumstances and the weaker the guarantee.

It is easy to see that an algorithm has more capabilities if it gains an extra oracle. The adversary $\mathcal{A}$ can have access to oracles for malicious adversaries (REG, CHALLINIT, CHALLRESP) and for semi-honest adversaries (REGHON, CHALLHONINIT, CHALLHONRESP). It is clear that the former oracles are stronger than the latter, yielding a stronger deniability guarantee.[8]

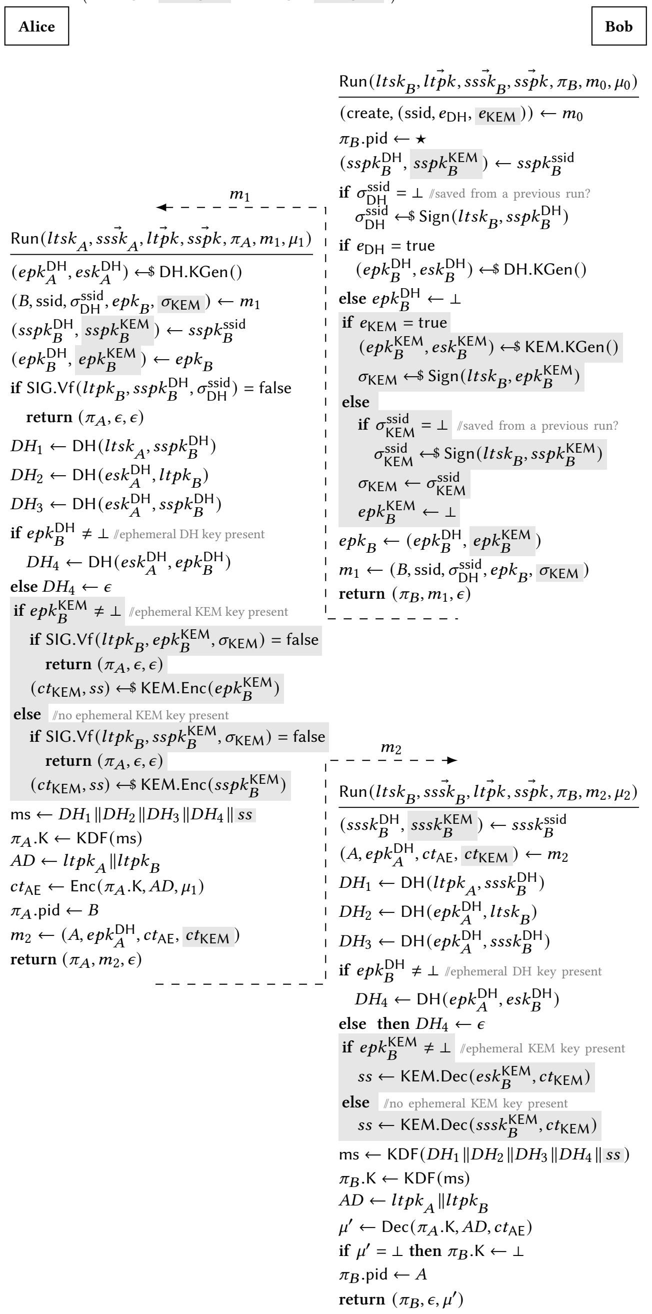
This leaves some relations open, e.g. how does deniability against semi-honest adversaries in the big brother model compare to deniability against malicious adversaries? While both notions remain incomparable, our model makes the differences explicit: We can argue about the "upper bound" (malicious adversaries in the big brother model) and the "lower bound" (semi-honest adversaries). Note that a deniability guarantee also holds for smaller numbers of users and semi-static keys: If it holds for n_p, n_{ss}, it also holds for $n'_p \leq n_p, n'_{ss} \leq n_{ss}$.[9]

4 DENIABILITY OF SIGNAL'S INITIAL HANDSHAKE

We look at Signal's initial handshake protocols—X3DH and its recent replacement PQXDH adding post-quantum confidentiality—and which deniability notions they fulfill. Note that both protocols aim for classical authentication and deniability, and not post-quantum deniability. Figure 3 gives an algorithmic overview of both protocols and the respective subsections give a textual description. The protocol specification [38, 43] uses only one long-term key per user, which is used for both the DH scheme and the signature scheme XEdDSA [47]. We follow [2] in treating them as two separate keys. We follow the protocol specification in using the session key directly as key for the AEAD scheme. Signal's implementation derives several values from the session key, among them the key and nonce used for the AEAD scheme. Beware that we consider Bob, who creates the pre-key bundle, as initiator, following e.g. [14, 54, 55] and contrasting [10].

We summarize our findings for both X3DH and PQXDH in Table 2, stated separately per role. For initiator (Bob) deniability, we have two results per adversary model to differentiate whether the Fake algorithm has access to a signature that the distinguisher does

```
KGenLT:
1 $(ltpk_U^{DH}, ltsk_U^{DH}) \leftarrow\$ \mathsf{DH.KGen}()$
2 $(ltpk_U^{SIG}, ltsk_U^{SIG}) \leftarrow\$ \mathsf{SIG.KGen}()$
3 return $\left((ltpk_U^{DH}, ltpk_U^{SIG}), (ltsk_U^{DH}, ltsk_U^{SIG})\right)$
KGenSS:
4 $(sspk_U^{DH}, sssk_U^{DH}) \leftarrow\$ \mathsf{DH.KGen}()$
5 $(sspk_U^{KEM}, sssk_U^{KEM}) \leftarrow\$ \mathsf{KEM.KGen}()$
6 return $\left((sspk_U^{DH}, sspk_U^{KEM}), (sssk_U^{DH}, sssk_U^{KEM})\right)$
```



Alice

Bob

```
Bob: $\mathsf{Run}(ltsk_B, \vec{ltpk}, \vec{sssk}_B, \vec{sspk}, \pi_B, m_0, \mu_0)$
$(\mathsf{create}, (\mathsf{ssid}, e_{DH}, e_{KEM})) \leftarrow m_0$
$\pi_B.\mathsf{pid} \leftarrow \star$
$(sspk_B^{DH}, sspk_B^{KEM}) \leftarrow sspk_B^{ssid}$
if $\sigma_{DH}^{ssid} = \perp$ //saved from a previous run?
  $\sigma_{DH}^{ssid} \leftarrow\$ \mathsf{Sign}(ltsk_B, sspk_B^{DH})$
if $e_{DH} = \mathsf{true}$
  $(epk_B^{DH}, esk_B^{DH}) \leftarrow\$ \mathsf{DH.KGen}()$
else $epk_B^{DH} \leftarrow \perp$
if $e_{KEM} = \mathsf{true}$
  $(epk_B^{KEM}, esk_B^{KEM}) \leftarrow\$ \mathsf{KEM.KGen}()$
  $\sigma_{KEM} \leftarrow\$ \mathsf{Sign}(ltsk_B, epk_B^{KEM})$
else
  if $\sigma_{KEM}^{ssid} = \perp$ //saved from a previous run?
    $\sigma_{KEM}^{ssid} \leftarrow\$ \mathsf{Sign}(ltsk_B, sspk_B^{KEM})$
  $\sigma_{KEM} \leftarrow \sigma_{KEM}^{ssid}$
  $epk_B^{KEM} \leftarrow \perp$
$epk_B \leftarrow (epk_B^{DH}, epk_B^{KEM})$
$m_1 \leftarrow (B, \mathsf{ssid}, \sigma_{DH}^{ssid}, epk_B, \sigma_{KEM})$
return $(\pi_B, m_1, \epsilon)$
```

m_1

```
Alice: $\mathsf{Run}(ltsk_A, \vec{sssk}_A, \vec{ltpk}, \vec{sspk}, \pi_A, m_1, \mu_1)$
$(epk_A^{DH}, esk_A^{DH}) \leftarrow\$ \mathsf{DH.KGen}()$
$(B, \mathsf{ssid}, \sigma_{DH}^{ssid}, epk_B, \sigma_{KEM}) \leftarrow m_1$
$(sspk_B^{DH}, sspk_B^{KEM}) \leftarrow sspk_B^{ssid}$
$(epk_B^{DH}, epk_B^{KEM}) \leftarrow epk_B$
if $\mathsf{SIG.Vf}(ltpk_B, sspk_B^{DH}, \sigma_{DH}^{ssid}) = \mathsf{false}$
  return $(\pi_A, \epsilon, \epsilon)$
$DH_1 \leftarrow \mathsf{DH}(ltsk_A, sspk_B^{DH})$
$DH_2 \leftarrow \mathsf{DH}(esk_A^{DH}, ltpk_B)$
$DH_3 \leftarrow \mathsf{DH}(esk_A^{DH}, sspk_B^{DH})$
if $epk_B^{DH} \neq \perp$ //ephemeral DH key present
  $DH_4 \leftarrow \mathsf{DH}(esk_A^{DH}, epk_B^{DH})$
else $DH_4 \leftarrow \epsilon$
if $epk_B^{KEM} \neq \perp$ //ephemeral KEM key present
  if $\mathsf{SIG.Vf}(ltpk_B, epk_B^{KEM}, \sigma_{KEM}) = \mathsf{false}$
    return $(\pi_A, \epsilon, \epsilon)$
  $(ct_{KEM}, ss) \leftarrow\$ \mathsf{KEM.Enc}(epk_B^{KEM})$
else //no ephemeral KEM key present
  if $\mathsf{SIG.Vf}(ltpk_B, sspk_B^{KEM}, \sigma_{KEM}) = \mathsf{false}$
    return $(\pi_A, \epsilon, \epsilon)$
  $(ct_{KEM}, ss) \leftarrow\$ \mathsf{KEM.Enc}(sspk_B^{KEM})$
$\mathsf{ms} \leftarrow DH_1 \| DH_2 \| DH_3 \| DH_4 \| ss$
$\pi_A.\mathsf{K} \leftarrow \mathsf{KDF}(\mathsf{ms})$
$AD \leftarrow ltpk_A \| ltpk_B$
$ct_{AE} \leftarrow \mathsf{Enc}(\pi_A.\mathsf{K}, AD, \mu_1)$
$\pi_A.\mathsf{pid} \leftarrow B$
$m_2 \leftarrow (A, epk_A^{DH}, ct_{AE}, ct_{KEM})$
return $(\pi_A, m_2, \epsilon)$
```

m_2

```
Bob: $\mathsf{Run}(ltsk_B, \vec{sssk}_B, \vec{ltpk}, \vec{sspk}, \pi_B, m_2, \mu_2)$
$(sssk_B^{DH}, sssk_B^{KEM}) \leftarrow sssk_B^{ssid}$
$(A, epk_A^{DH}, ct_{AE}, ct_{KEM}) \leftarrow m_2$
$DH_1 \leftarrow \mathsf{DH}(ltpk_A, sssk_B^{DH})$
$DH_2 \leftarrow \mathsf{DH}(epk_A^{DH}, ltsk_B)$
$DH_3 \leftarrow \mathsf{DH}(epk_A^{DH}, sssk_B^{DH})$
if $epk_B^{DH} \neq \perp$ //ephemeral DH key present
  $DH_4 \leftarrow \mathsf{DH}(epk_A^{DH}, esk_B^{DH})$
else then $DH_4 \leftarrow \epsilon$
if $epk_B^{KEM} \neq \perp$ //ephemeral KEM key present
  $ss \leftarrow \mathsf{KEM.Dec}(esk_B^{KEM}, ct_{KEM})$
else //no ephemeral KEM key present
  $ss \leftarrow \mathsf{KEM.Dec}(sssk_B^{KEM}, ct_{KEM})$
$\mathsf{ms} \leftarrow \mathsf{KDF}(DH_1 \| DH_2 \| DH_3 \| DH_4 \| ss)$
$\pi_B.\mathsf{K} \leftarrow \mathsf{KDF}(\mathsf{ms})$
$AD \leftarrow ltpk_A \| ltpk_B$
$\mu' \leftarrow \mathsf{Dec}(\pi_A.\mathsf{K}, AD, ct_{AE})$
if $\mu' = \perp$ then $\pi_B.\mathsf{K} \leftarrow \perp$
$\pi_B.\mathsf{pid} \leftarrow A$
return $(\pi_B, \epsilon, \mu')$
```

Figure 3: Signal's initial handshake protocols X3DH **and** PQXDH**. The** KEM **with gray background is exclusive to** PQXDH**.**

[8]Assuming that the CHALLINIT, CHALLRESP oracles come with an INIT oracle, since they are otherwise useless.
[9]A reduction can simply withhold some keys from the inner adversary.

not know, i.e. a *private signature*, or if Fake and the distinguisher know the same signatures, i.e. *public signatures*. For X3DH either case works (see Remark 4.1), while for PQXDH we *must have* a private signature for the ephemeral KEM key (see Theorem 4.4 and Remark 4.4). Without further restrictions, we achieve 1-out-of-2 deniability against semi-honest adversaries. We use a knowledge of DH assumption (see Definitions 2.1 and B.4) against malicious adversaries and additionally plaintext awareness of the KEM (see Definition 2.2) for PQXDH to simulate the session key. Concerning responder (Alice) deniability, PQXDH matches the guarantees of X3DH without limitations. For malicious adversaries in the big brother model, i.e. the strongest possible adversary model, we give our thoughts in Remarks 4.2 and 4.5. Note that Table 2 implies all other notions: Each theorem still holds if you weaken the adversary, weaken the distinguisher, add an oracle to O_F, or add auxiliary info (without the Fake algorithm using it).

While our analysis relies on the random oracle model, we do not program the RO, following [23] and in line with the results of [46]. All Fake algorithms abort on receiving malformed inputs. We state the expected message format explicitly in the pseudocode.

4.1 X3DH

We recall how Signal's classical initial handshake protocol X3DH [43] works, see Figure 3. First, Bob signs his semi-static public key (or retrieves a previously created signature) under his long-term key. Bob samples an ephemeral DH key pair if the boolean e_{DH} is set. Bob's three public keys (long-term, semi-static with id ssid, and optionally ephemeral) and the signature are Bob's first message. We call this first message Bob's *pre-key bundle*, which he sends to the key server and is not tied to any particular peer. Second, Alice parses the message, verifies the signature, samples an ephemeral DH key herself, and computes four DH shared secrets (or three if Bob's message does not include an ephemeral key) between her keys and Bob's keys (long-term–semi-static, ephemeral–long-term, ephemeral–semi-static, ephemeral–ephemeral), and finally derives the session key from the DH shared secrets. She encrypts her first user message with the AEAD scheme under the session key[10] and sends this AEAD ciphertext and her ephemeral public key back to Bob. Third, Bob computes the DH shared secrets and derives the session key in the same way. He tries to decrypt the AEAD ciphertext and rejects the session key if decryption fails.

To simulate a transcript and session key, we face two challenges: obtaining all DH shared secrets and obtaining a valid signature on Bob's semi-static key. We briefly discuss our approaches for each of the two challenges, as well as the difficulties arising with big brother distinguishers, before going into the actual proofs.

4.1.1 Obtaining the DH shared secrets. For adversaries that create keys honestly, the Fake algorithm knows the necessary DH secret keys to compute all DH shared secrets itself (via the SK oracle or the semi-honest peer's randomness). Against malicious adversaries, the Fake algorithm uses the knowledge of DH extractor (of the EKDH assumption, see Definition 2.1) to extract the remaining DH shared secrets from the adversary, following [56]. By assumption, if the extractor fails, no other extractor can succeed in a given time with more than a given probability. This allows Fake to set a random session key in case the extraction fails.

[10]In practice, Signal uses an elaborate key scheduling algorithm. The actual key and nonce used for encryption are deterministically derived from the session key.

4.1.2 Obtaining a signature on Bob's semi-static key. Recall that the semi-static key can be used for several sessions, so neither the semi-static key nor the signature on the semi-static key are tied to a particular session. The Fake algorithm has two options (shown in Figure 4) to obtain a signature on Bob's semi-static key: First, from a pre-key bundle in the auxiliary info (the pre-key bundle may be part of a complete transcript), as previously done by [56]. Second, Fake may query the $\textsc{User}_{\text{initiator},\perp}$ oracle (i.e. the oracle acts as initiator) to get a pre-key bundle including the signature on the semi-static key. Note that the Fake algorithm may use the SK oracle to learn the peer's secret key but does not have an option to learn the secret key of the session owner.

Both options rely on Fake obtaining a signature from the pre-key bundle. The first option (using aux) models that the adversary is able to eavesdrop on previous protocol executions and, hence, Fake can use aux as well. The second option (the $\textsc{User}_{\text{initiator},\perp}$ oracle) models that Fake may start an independent session with the victim, i.e. anybody who can start a session with the victim can produce a transcript and session key. If Fake uses aux and aux is known to the distinguisher, then we have *public signatures*. Otherwise, we have *private signatures*, i.e. Fake knows signatures that the distinguisher does not. For X3DH either setting works. Looking ahead, this will make a difference for PQXDH.

4.1.3 Big Brother distinguishers. Ideally, we want to have deniability results against malicious adversaries in the big brother model. Against malicious adversaries, we rely on the knowledge of DH assumption, which we cannot apply in the big brother model. We discuss the challenges to show responder deniability in the big brother model in Remark 4.2.

4.1.4 Theorems for X3DH. Here we give our results for initiator deniability of X3DH. We defer the proofs to Appendix D. For responder deniability, we show the same results for both X3DH and PQXDH. We state them in Section 4.2.3 (Theorems 4.8 and 4.9).

We show that Bob can deny (*initiator deniability*) participation in a protocol run against a semi-honest Alice (*against semi-honest adversaries*, Thm. 4.1), against a malicious Alice that honestly generated her keys (*against malicious adversaries with honest keys*, Thm. 4.2), and against a malicious Alice (*against malicious adversaries*, Thm. 4.3). He does so by showing that Alice (1-*out-of*-2, i.e. anybody using Alice's secret keys, Thms. 4.1 and 4.2) or anybody (1-*out-of*-∞, Thm. 4.3) could have produced the transcript and session key herself, assuming she has previously observed a valid transcript with Bob as initiator (AuxPrep *yielding a valid pre-key bundle per* ssid *and user*). This holds even against a big brother distinguisher (*in the big brother model*, Thm. 4.1), who shares Alice's observations (aux *known to the distinguisher*, Thms. 4.1 to 4.3).

Theorem 4.1. *The* X3DH *protocol as shown in Figure 3 is* $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* $(\{\textsc{RegHon}, \textsc{ChallHonInit}\}, \{SK\}, \{SKs, \textsc{aux}\})$*-oracles and auxiliary info sampled with* AuxPrep *yielding a valid pre-key bundle per* ssid *and user, and* aux *known to the distinguisher, where* n_p *is the number of parties and* n_{ss} *the number of semi-static keys per party, and* $q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}$ *are arbitrary and* $\epsilon_{\mathcal{D}} = 0$.

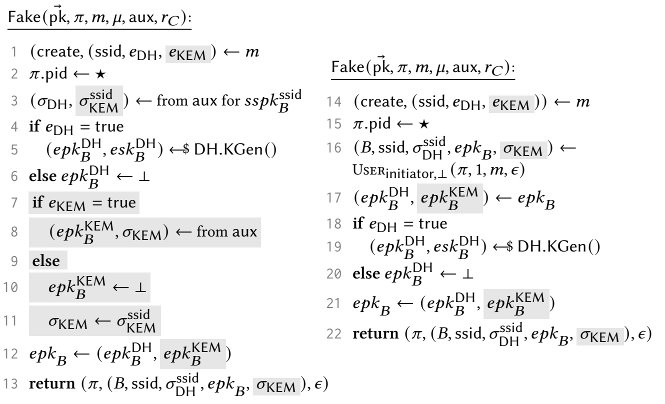
X3DH	PQXDH	σ	Assumption
semi-honest adv., big brother dist., 1-out-of-2			
✓ Th. 4.1	✗ Th. 4.4	👁	-
✓ Rm. 4.1	✓ Th. 4.5	👁̸	-
malicious adversaries with honest keys, 1-out-of-2			
✓ Th. 4.2	✗ Rm. 4.4	👁	K2DHA, KDF RO
✓ Rm. 4.1	✓ Th. 4.6	👁̸	K2DHA, KDF RO, KEM PA1
malicious adversaries, 1-out-of-∞			
✓ Th. 4.3	✗ Rm. 4.4	👁	EKDHA, KDF RO
✓ Rm. 4.1	✓ Th. 4.7	👁̸	EKDHA, KDF RO, KEM PA1
malicious adv, big brother dist., 1-out-of-∞			
● Rm. 4.2	● Rm. 4.5	👁̸	-

(a) Initiator (Bob) deniability

X3DH	PQXDH	σ	Assumption
malicious adv. w/ honest keys, big brother dist., 1-out-of-2			
✓ Th. 4.8		$\perp$	-
malicious adversaries, 1-out-of-∞			
✓ Th. 4.9		$\perp$	EKDHA, KDF RO
malicious adversaries, big brother distinguisher, 1-out-of-∞			
● Rm. 4.2	● Rm. 4.5	$\perp$	-

(b) Responder (Alice) deniability

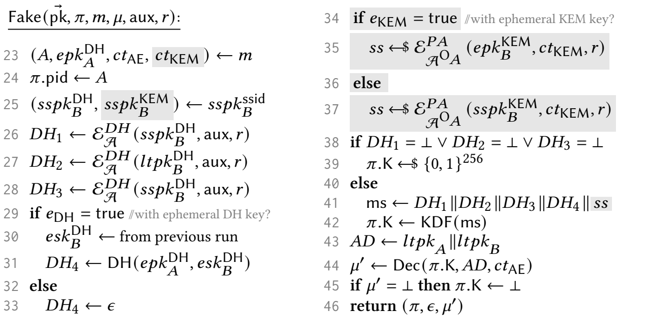
Table 2: Comparing deniability of X3DH **and** PQXDH**: Theorems (and Remarks) in the first two columns address the model in the continuous line above (indicating adversarial capabilities, big brother model, and how many people can fake a transcript, i.e. whether SK $\in$ O$_F$), the third column indicates if** Fake **requires a signature that the distinguisher is not aware of with 👁̸ (USER$_{\text{initiator},y}$ $\in$ O$_F$ or both** AuxPrep **yields appropriate signatures and AUX $\notin$ O$_D$) or a signature that the distinguisher may learn with 👁 (USER$_{\text{initiator},y}$ $\in$ O$_F$ or** AuxPrep **yields appropriate signatures) or no signature at all with $\perp$, and the final column states the assumptions needed. We denote that a property holds by ✓, does not hold by ✗, or that we currently do not have a proof by ●. For the knowledge of DH assumptions (K2DH, EKDHA), see Definitions 2.1 and B.4; the plaintext awareness assumption on** KEM **(gray background , see Definition 2.2) only applies to** PQXDH**.**



```
Fake($\vec{\mathrm{pk}}, \pi, m, \mu, \mathrm{aux}, r_C$):
(create, (ssid, $e_{\mathrm{DH}}$, $e_{\mathrm{KEM}}$) $\leftarrow m$
$\pi$.pid $\leftarrow \star$
($\sigma_{\mathrm{DH}}$, $\sigma^{\mathrm{ssid}}_{\mathrm{KEM}}$) $\leftarrow$ from aux for $sspk^{\mathrm{ssid}}_B$
if $e_{\mathrm{DH}}$ = true
    $(epk^{\mathrm{DH}}_B, esk^{\mathrm{DH}}_B) \leftarrow\!\$\ \mathrm{DH.KGen}()$
else $epk^{\mathrm{DH}}_B \leftarrow \perp$
if $e_{\mathrm{KEM}}$ = true
    $(epk^{\mathrm{KEM}}_B, \sigma_{\mathrm{KEM}}) \leftarrow$ from aux
else
    $epk^{\mathrm{KEM}}_B \leftarrow \perp$
    $\sigma_{\mathrm{KEM}} \leftarrow \sigma^{\mathrm{ssid}}_{\mathrm{KEM}}$
$epk_B \leftarrow (epk^{\mathrm{DH}}_B, epk^{\mathrm{KEM}}_B)$
return $(\pi, (B, \mathrm{ssid}, \sigma^{\mathrm{ssid}}_{\mathrm{DH}}, epk_B, \sigma_{\mathrm{KEM}}), \epsilon)$
```

```
Fake($\vec{\mathrm{pk}}, \pi, m, \mu, \mathrm{aux}, r_C$):
(create, (ssid, $e_{\mathrm{DH}}$, $e_{\mathrm{KEM}}$)) $\leftarrow m$
$\pi$.pid $\leftarrow \star$
$(B, \mathrm{ssid}, \sigma^{\mathrm{ssid}}_{\mathrm{DH}}, epk_B, \sigma_{\mathrm{KEM}}) \leftarrow$
USER$_{\mathrm{initiator},\perp}(\pi, 1, m, \epsilon)$
$(epk^{\mathrm{DH}}_B, epk^{\mathrm{KEM}}_B) \leftarrow epk_B$
if $e_{\mathrm{DH}}$ = true
    $(epk^{\mathrm{DH}}_B, esk^{\mathrm{DH}}_B) \leftarrow\!\$\ \mathrm{DH.KGen}()$
else $epk^{\mathrm{DH}}_B \leftarrow \perp$
$epk_B \leftarrow (epk^{\mathrm{DH}}_B, epk^{\mathrm{KEM}}_B)$
return $(\pi, (B, \mathrm{ssid}, \sigma^{\mathrm{ssid}}_{\mathrm{DH}}, epk_B, \sigma_{\mathrm{KEM}}), \epsilon)$
```

Figure 4: The Fake **algorithms simulating Bob's pre-key bundle (initiator deniability) in Theorems 4.1 to 4.3 and 4.5 to 4.7.**



```
Fake($\vec{\mathrm{pk}}, \pi, m, \mu, \mathrm{aux}, r$):
$(A, epk^{\mathrm{DH}}_A, ct_{\mathrm{AE}}, ct_{\mathrm{KEM}}) \leftarrow m$
$\pi$.pid $\leftarrow A$
$(sspk^{\mathrm{DH}}_B, sspk^{\mathrm{KEM}}_B) \leftarrow sspk^{\mathrm{ssid}}_B$
$DH_1 \leftarrow \mathcal{E}^{DH}_{\mathcal{A}}(sspk^{\mathrm{DH}}_B, \mathrm{aux}, r)$
$DH_2 \leftarrow \mathcal{E}^{DH}_{\mathcal{A}}(ltpk^{\mathrm{DH}}_B, \mathrm{aux}, r)$
$DH_3 \leftarrow \mathcal{E}^{DH}_{\mathcal{A}}(sspk^{\mathrm{DH}}_B, \mathrm{aux}, r)$
if $e_{\mathrm{DH}}$ = true //with ephemeral DH key?
    $esk^{\mathrm{DH}}_B \leftarrow$ from previous run
    $DH_4 \leftarrow \mathrm{DH}(epk^{\mathrm{DH}}_A, esk^{\mathrm{DH}}_B)$
else
    $DH_4 \leftarrow \epsilon$
if $e_{\mathrm{KEM}}$ = true //with ephemeral KEM key?
    $ss \leftarrow\!\$\ \mathcal{E}^{PA}_{\mathcal{A}^{\mathrm{O}_A}}(epk^{\mathrm{KEM}}_B, ct_{\mathrm{KEM}}, r)$
else
    $ss \leftarrow\!\$\ \mathcal{E}^{PA}_{\mathcal{A}^{\mathrm{O}_A}}(sspk^{\mathrm{KEM}}_B, ct_{\mathrm{KEM}}, r)$
if $DH_1 = \perp \vee DH_2 = \perp \vee DH_3 = \perp$
    $\pi.\mathrm{K} \leftarrow\!\$\ \{0,1\}^{256}$
else
    $\mathrm{ms} \leftarrow DH_1 \| DH_2 \| DH_3 \| DH_4 \| ss$
    $\pi.\mathrm{K} \leftarrow \mathrm{KDF}(\mathrm{ms})$
$AD \leftarrow ltpk_A \| ltpk_B$
$\mu' \leftarrow \mathrm{Dec}(\pi.\mathrm{K}, AD, ct_{\mathrm{AE}})$
if $\mu' = \perp$ then $\pi.\mathrm{K} \leftarrow \perp$
return $(\pi, \epsilon, \mu')$
```

Figure 5: The Fake **algorithm processing Alice's message (initiator deniability) in Theorems 4.3 and 4.7.**

REMARK 4.1 (OBTAINING THE SIGNATURE). *In Theorems 4.1 to 4.3, the* Fake *algorithm obtains the signature on the semi-static DH key from the auxiliary info* aux *(cp. the left-hand side of Figure 4). Instead, the* Fake *algorithm may also learn the signature via the* USER$_{\mathrm{initiator},\perp}$ *oracle (cp. the right-hand side of Figure 4), resulting in* partial deniability wrt. a USER$_{\mathrm{initiator},\perp}$ oracle. *Hence, these theorems also hold for* AuxPrep $= \perp$ *if* O_F *additionally includes* USER$_{\mathrm{initiator},\perp}$.

The following two theorems use knowledge of DH assumptions, which [56] have introduced to show deniability of X3DH in their model. We discuss the differences between their proof and our proofs in Appendices D.2 and D.3.

THEOREM 4.2. *If the* $(t_{\mathcal{A}DH}, t_{\mathcal{E}DH}, t_{\mathcal{D}DH}, \epsilon_{\mathcal{D}DH})$*-Knowledge of 2DH (K2DH) Assumption holds and* KDF *is a random oracle, then the* X3DH *protocol as shown in Figure 3 is* $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* ({REGHON, INIT, CHALLINIT}, {SK}, {AUX})*-oracles and auxiliary inputs sampled with* AuxPrep *yielding a valid pre-key bundle per* ssid *and user, and* aux *known to the distinguisher, where* n_p *is the number of parties and* n_{ss} *the number of semi-static keys per party,* $q_{O_A}, q_{O_F}, q_{O_D}$ *are arbitrary, and* $t_{\mathcal{A}} \approx t_{\mathcal{A}DH}$, $t_{\mathcal{D}} \approx t_{\mathcal{D}DH}$*, and* $\epsilon_{\mathcal{D}} \leq \epsilon_{\mathcal{D}DH}$.

THEOREM 4.3. *If the* $(t_{\mathcal{A}DH}, t_{\mathcal{E}DH}, t_{\mathcal{D}DH}, \epsilon_{\mathcal{D}DH})$*-Extended Knowledge of DH (EKDH) Assumption holds and* KDF *is a random oracle, then the* X3DH *protocol as shown in Figure 3 is* $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* ({REG, INIT, CHALLINIT}, $\emptyset$, {AUX})*-oracles and auxiliary info sampled with* AuxPrep *yielding a valid pre-key bundle per* ssid *and user, and* aux *known to the distinguisher, where* n_p *is the number of parties and* n_{ss} *the number of semi-static keys per party,* $q_{O_A}, q_{O_F}, q_{O_D}$ *are arbitrary, and* $t_{\mathcal{A}} \approx 3 \cdot t_{\mathcal{A}DH}$, $t_{\mathcal{D}} \approx t_{\mathcal{D}DH}$*, and* $\epsilon_{\mathcal{D}} \leq \epsilon_{\mathcal{D}DH}$.

REMARK 4.2 (DENIABILITY AGAINST MALICIOUS ADVERSARIES IN THE BIG BROTHER MODEL). *In the big brother setting, we cannot* *apply the EKDH assumption (since the distinguisher gets extra inputs compared to the adversary and* Fake*). Hence, if the extractor fails a big brother distinguisher may be able to compute the correct session key with probability* $> \epsilon_{\mathcal{D}DH}$*. Thereby, the big brother distinguisher can tell the two cases apart. We discuss another (unsuccessful) strategy in Section 5.*

4.2 PQXDH

We review PQXDH [38] with the help of Figure 3. PQXDH extends X3DH by including a KEM to achieve post-quantum confidentiality. In particular, Bob's pre-key bundle includes an ephemeral KEM key (if the boolean e_{KEM} is set) or a semi-static KEM key (otherwise), and a signature on this KEM public key. Alice verifies the signature on the KEM key (and on the semi-static DH key as before) and encapsulates against Bob's KEM public key. Alice's message consists of her ephemeral DH key, the AEAD ciphertext, and the KEM ciphertext. Both parties compute the session key from the three or four DH shared secrets and the KEM shared secret, and use the session key for encryption and decryption of the AEAD ciphertext, respectively.

Note that in practice, Bob's pre-key bundle includes a semi-static KEM key only if no ephemeral KEM key is used. We model this by not using the semi-static KEM key if an ephemeral KEM key exists; though we formally unwrap it from the semi-static key.

Since PQXDH builds on top of X3DH, simulating a transcript bears the same difficulties as for X3DH with some new ones added on top: obtaining a signature for the KEM key (which may be ephemeral or semi-static) and the KEM shared secret.

4.2.1 Obtaining a signature on Bob's semi-static and ephemeral keys. We need a signature on Bob's semi-static DH key for X3DH, and also on Bob's ephemeral or semi-static KEM key for PQXDH. So we need to cover signatures on semi-static *and* ephemeral keys.

We have the same two options as for X3DH to obtain a signature: from aux and from the $\textsc{User}_{\mathsf{initiator},y}$ oracle, see Figure 4. If Fake uses aux and aux is known to the distinguisher, then we have *public signatures*. Otherwise, we have *private signatures*, i.e. Fake knows signatures that the distinguisher does not. The distinguisher can detect reuse of public signatures, see Theorem 4.4. Hence, we require private signatures, see Remark 4.4.

4.2.2 Obtaining the KEM shared secret. For responder deniability, Fake can learn the KEM shared secret by encapsulating against the KEM public key. For initiator deniability against malicious adversaries, we rely on plaintext awareness of the KEM (see Definition 2.2), following [30].[11]

4.2.3 Theorems for PQXDH*.* Here we give our results for initiator deniability of PQXDH and or responder deniability of both X3DH and PQXDH. We defer the proofs to Appendix E.We start with initiator deniability and show that faking a transcript requires a signature that the distinguisher is not aware of, i.e. a private signature. In particular, we show that Bob cannot deny (*initiator deniability*) participation in a protocol run against a semi-honest Alice (*against semi-honest adversaries*). This follows from nobody, not even Alice (1-*out-of*-2, i.e. using Alice's secret keys) being able to produce the transcript and session key herself, even assuming she has previously observed a valid transcript with Bob as initiator (AuxPrep *yielding a valid pre-key bundle per* ssid *and user*). This holds even against a big brother distinguisher (*in the big brother model*), who shares Alice's observations (aux *known to the distinguisher*).

An indistinguishable transcript contains a valid signature on the ephemeral KEM key and the distinguisher must not have seen that signature before. Since the Fake algorithm cannot get this signature from anywhere, it must be a forgery of the signature scheme.

Theorem 4.4. *If* SIG *is* $(q_{\mathsf{SIG}}, t_{\mathsf{SIG}}, \epsilon_{\mathsf{SIG}})$*-unforgeable,* KEM *has a key collision probability bounded by* γ_{coll}*, and the public key spaces of* KEM *and* DH *are disjoint, then the* PQXDH *protocol as shown in Figure 3 is* not $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* $(\{\textsc{RegHon}, \textsc{ChallHonInit}\}, \{SK\}, \{SKs, \textsc{aux}\})$*-oracles and auxiliary inputs sampled with* AuxPrep *yielding a valid pre-key bundle per* ssid *and user and* q_{CI} *pre-key bundles per user with ephemeral KEM keys, where* $n_p \geq 2$ *is the number of parties and* n_{ss} *the number of semi-static keys per party,* $q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}$ *are arbitrary,* $q_{\mathsf{SIG}} = 2 \cdot n_{ss} + q_{CI}$*,* $t_{\mathsf{SIG}} \approx t_{\mathcal{A}}$*, and* $\epsilon_{\mathcal{D}} \not< 1 - n_{ss} \cdot \gamma_{coll} - \epsilon_{\mathsf{SIG}}$*.*

Proof of Theorem 4.4. For contradiction, we assume PQXDH was $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$-deniable and build a reduction to $(q_{\mathsf{SIG}}, t_{\mathsf{SIG}}, \epsilon_{\mathsf{SIG}})$-unforgeability of SIG. Note that the reduction relies on the Fake algorithm to win its unforgeability game. As such, the reduction simulates both the deniability game and the deniability adversary around the Fake algorithm. The reduction receives a challenge key $\mathsf{pk}_{\mathsf{SIG}}$ as input and proceeds to run $\mathsf{KeyPrep}(n_p, n_{ss})$ to generate key pairs for all users. Wlog. we determine two distinct users A and B. The reduction saves the long-term signing key pair of B as $(ltpk_B^{\mathsf{SIG}}, ltsk_B^{\mathsf{SIG}})$, replaces the public key with the challenge key $\mathsf{pk}_{\mathsf{SIG}}$, and saves the resulting list of public keys as $\vec{\mathsf{pk}}$. The reduction prepares the auxiliary info aux according to AuxPrep; to produce the pre-key bundles for B, the reduction queries its $\textsc{Sign}$ oracle to obtain signatures under $\mathsf{pk}_{\mathsf{SIG}}$ for all of B's semi-static DH keys and KEM keys and for q_{CI} ephemeral KEM keys, totaling $2 \cdot n_{ss} + q_{CI}$ queries. Since the reduction controls both the (deniability) challenger and the (deniability) adversary, it acts as if the adversary challenges B to a full handshake with A[12] while $b = 1$. The Fake algorithm has to answer for this query. If Fake queries its SK oracle, the reduction replies with $(ltsk_A, \vec{sssk}_A)$. The Fake algorithm returns a protocol message containing a signed ephemeral KEM key for B. The reduction returns this ephemeral KEM key and signature tuple $(epk_B^{\mathsf{KEM}}, \sigma_{\mathsf{KEM}})$.

We include the running time of the oracles $O_A = \{\textsc{RegHon}, \textsc{ChallHonInit}\}, O_F = \{SK\}$ in $t_{\mathcal{A}}$, allowing for $t_{\mathsf{SIG}} \approx t_{\mathcal{A}}$. The reduction simulates the game $\mathcal{G}^{O_A,O_F,O_D\text{-}den}_{\mathsf{PQXDH},\mathsf{AuxPrep},n_p,n_{ss}}(\mathcal{A}, \mathcal{D})$ around Fake faithfully, since it can mimic the behavior of the game and answer as the game would.

The distinguisher can tell if Fake uses a signature on the ephemeral key from aux since the distinguisher knows aux itself (via the $\textsc{aux}$ oracle). We exclude this by assuming PQXDH to be deniable. The probability of epk_B^{KEM} colliding with one of B's n_{ss} KEM keys signed in AuxPrep is $n_{ss} \cdot \gamma_{coll}$ (and 0 to collide with a DH public key). We need to exclude this probability since these signatures were created by the $\textsc{Sign}$ oracle of the reduction. If the transcript contained

[11]For initiator deniability against semi-honest adversaries, Fake can learn the randomness used for encapsulation and thereby obtain the KEM shared secret.

[12]via $\textsc{ChallHonInit}(B, A, (\text{create}, (\text{ssid} = 1, e_{\mathsf{DH}} = \text{true}, e_{\mathsf{KEM}} = \text{true})))$

an invalid signature, then the transcript could not be $(t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$-indistinguishable. Hence, the signature must be valid and on a fresh message. Therefore, the reduction wins if the distinguisher loses and we get $\epsilon_{\mathsf{SIG}} \geq 1 - n_{ss} \cdot \gamma_{coll} - \epsilon_{\mathcal{D}}$ and $\epsilon_{\mathcal{D}} \geq 1 - n_{ss} \cdot \gamma_{coll} - \epsilon_{\mathsf{SIG}}$, which is not close to $\frac{1}{2}$. □

Remark 4.3 (Domain separators). *If signatures on the KEM public keys include a domain separator for ephemeral and semi-static use, then we do not have the loss of $n_{ss} \cdot \gamma_{coll}$ in the above theorem.*

Remark 4.4 (Fake needs a private signature for initiator deniability). *In Theorem 4.4 the distinguisher detects* Fake *since* Fake *does not learn any "private" signatures on ephemeral keys that the distinguisher does not know. In Theorems 4.5 to 4.7, the* Fake *algorithm obtains signatures on the semi-static DH key and on the KEM key from the auxiliary info* aux *(cp. the left-hand side of Figure 4), while the distinguisher* does not, *i.e.* $\textsc{aux} \notin O_D$. *Alternatively, the* Fake *algorithm may also learn the signatures via the* $\textsc{User}_{\text{initiator},\perp}$ *oracle (cp. the right-hand side of Figure 4), resulting in* partial deniability wrt. a $\textsc{User}_{\text{initiator},\perp}$ oracle. *Hence, these theorems also hold for* AuxPrep $= \perp$ *if* O_F *additionally includes* $\textsc{User}_{\text{initiator},\perp}$. *In either case the distinguisher does not see the signature, i.e. the signature remains "private".*

Bob can deny (*initiator deniability*) participation in a protocol run against a semi-honest Alice (*against semi-honest adversaries*, Thm. 4.5), against a malicious Alice that honestly generated her keys (*against malicious adversaries with honest keys*, Thm. 4.6), and against a malicious Alice (*against malicious adversaries*, Thm. 4.7). He does so by showing that Alice (1-*out-of*-2, i.e. anybody using Alice's secret keys, Thms. 4.5 and 4.6) or anybody (1-*out-of*-∞, Thm. 4.7) could have produced the transcript and session key herself, assuming she has previously observed a valid transcript with Bob as initiator (AuxPrep *yielding a valid pre-key bundle per* ssid *and user and* q_{CI} *pre-key bundles per user with ephemeral KEM keys*, Thms. 4.5 to 4.7). This holds against big brother distinguishers (*in the big brother model*, Thm. 4.5) who do not learn Alice's observations (aux *unknown to the distinguisher*, Thms. 4.5 to 4.7).

Theorem 4.5. *The* PQXDH *protocol as shown in Figure 3 is* $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* $(\{\textsc{RegHon}, \textsc{ChallHonInit}\}, \{SK\}, \{SKs\})$*-oracles and auxiliary inputs sampled with* AuxPrep *yielding a valid pre-key bundle per* ssid *and user and* q_{CI} *pre-key bundles per user with ephemeral KEM keys, where* n_p *is the number of parties and* n_{ss} *the number of semi-static keys per party,* $q_{CI} \in q_{O_A}$, $q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}$ *are arbitrary, and* $\epsilon_{\mathcal{D}} = 0$.

For the next two theorems we need plaintext awareness.

Theorem 4.6. *If the* $(t_{\mathcal{A}^{DH}}, t_{\mathcal{E}^{DH}}, t_{\mathcal{D}^{DH}}, \epsilon_{\mathcal{D}^{DH}})$*-Knowledge of 2DH (K2DH) Assumption holds,* KEM *is* $(n_k, t_C, t_{\mathcal{E}^{PA}}, t_{\mathcal{D}^{PA}}, \epsilon_{\mathcal{D}^{PA}})$*-PA1 secure, and* KDF *is a random oracle, then the* PQXDH *protocol as shown in Figure 3 is* $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* $(\{\textsc{RegHon}, \textsc{Init}, \textsc{ChallInit}\}, \{SK\}, \emptyset)$*-oracles and auxiliary inputs sampled with* AuxPrep *yielding a valid pre-key bundle per* ssid *and user and* q_{CI} *pre-key bundles per user with ephemeral KEM keys, where* n_p *is the number of parties and* n_{ss} *the number of semi-static keys per party, the number of keys for the PA assumption is* $n_k = n_p \cdot n_{ss} + q_{CI}$, $q_{CI} \in q_{O_A}$, $q_{O_A}, q_{O_F}, q_{O_D}$ *are arbitrary,* $t_{\mathcal{A}} \approx t_C \approx t_{\mathcal{E}^{DH}}$, $t_{\mathcal{D}} \approx t_{\mathcal{D}^{DH}} \approx t_{\mathcal{D}^{PA}}$*, and* $\epsilon_{\mathcal{D}} \leq \epsilon_{\mathcal{D}^{DH}} + \epsilon_{\mathcal{D}^{PA}}$.

Theorem 4.7. *If the* $(t_{\mathcal{A}^{DH}}, t_{\mathcal{E}^{DH}}, t_{\mathcal{D}^{DH}}, \epsilon_{\mathcal{D}^{DH}})$*-Extended Knowledge of DH (EKDH) Assumption holds,* KEM *is* $(n_k, t_C, t_{\mathcal{E}^{PA}}, t_{\mathcal{D}^{PA}}, \epsilon_{\mathcal{D}^{PA}})$*-PA1 secure, and* KDF *is a random oracle, then the* PQXDH *protocol as shown in Figure 3 is* $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* $(\{\textsc{Reg}, \textsc{Init}, \textsc{ChallInit}\}, \emptyset, \emptyset)$*-oracles and auxiliary inputs sampled with* AuxPrep *yielding a valid pre-key bundle per* ssid *and user and* q_{CI} *pre-key bundles per user with ephemeral KEM keys, where* n_p *is the number of parties and* n_{ss} *the number of semi-static keys per party,* $q_{CI} \in q_{O_A}$, $q_{O_A}, q_{O_F}, q_{O_D}$ *are arbitrary, and* $t_{\mathcal{A}} \approx t_C \approx 3 \cdot t_{\mathcal{E}^{DH}}$, $t_{\mathcal{D}} \approx t_{\mathcal{D}^{DH}} \approx t_{\mathcal{D}^{PA}}$*, and* $\epsilon_{\mathcal{D}} \leq \epsilon_{\mathcal{D}^{DH}} + \epsilon_{\mathcal{D}^{PA}}$.

Proof of Theorem 4.7. We give the Fake algorithm in Figure 4 on the right-hand side and Figure 5.

The Fake algorithm reuses the pre-key bundle obtained from the $\textsc{User}_{\text{initiator},\perp}$ oracle but replaces the ephemeral DH key with a freshly sampled key (if mandated by e_{DH}). Furthermore, to process Alice's message, the Fake algorithm requires several techniques: The Fake algorithm relies on the extractor from the EKDH assumption to learn DH_1 through DH_3 (lines 26, 27, and 28) and uses Bob's ephemeral key (from its previous run) to compute DH_4. Additionally, it utilizes the plaintext extractor for the KEM to learn the KEM shared secret (line 35 or 37). If any of the three extractions under the EKDH assumption fail, Fake sets a random key as session key. Else, it computes the session key honestly. It decrypts the AEAD ciphertext and aborts if decryption fails.

The Fake algorithm creates Bob's pre-key bundle virtually identical to Run. We argue via a series of game hops that the distinguisher cannot tell whether Alice's message was processed by Run or Fake.

Game 0 = Run. The initial game is the original deniability game $\mathcal{G}^{\mathsf{O}_A,\mathsf{O}_F,\mathsf{O}_D\text{-}den}_{\mathsf{PQXDH},\mathsf{AuxPrep},n_p,n_{ss}}(\mathcal{A}, \mathcal{D})$ with challenge $b = 0$, executing Run.

Game 1 (PA1 extractor). We replace the honest decapsulation of the KEM ciphertext with the PA extractor in lines 35 and 37. (In each invocation of Fake for processing Alice's message, exactly one of these two lines is executed, depending on e_{DH}.)

We bound the advantage difference introduced by this step by the $(n_k, t_C, t_{\mathcal{E}^{PA}}, t_{\mathcal{D}^{PA}}, \epsilon_{\mathcal{D}^{PA}})$-PA1 assumption. We consider the reduction $\mathcal{B}$—running the adversary with its oracles, excluding calls to the $\textsc{Decaps}$ oracle—as ciphertext creator C of the PA1 game. The reduction $\mathcal{B}$ is started with a list of KEM keys $\vec{\mathsf{pk}}_{\mathsf{KEM}}$ and explicit randomness $r_C = (r_{\mathcal{A}}, r_R, r_C)$.

The reduction samples $n_p \cdot n_{ss}$ DH keys and sorts them by user into $\vec{\mathsf{pk}}$. Next, it prepares the auxiliary info aux, which consists of one pre-key bundle per user and semi-static key. Note that for both sampling the DH keys and preparing the auxiliary info, the reduction uses hard-wired randomness. (Without hard-wiring the randomness, the PA1 extractor would require this randomness as input since we view the reduction as ciphertext creator.) Then, it starts the adversary $\mathcal{A}$ on the auxiliary info aux and the list of public keys $\vec{\mathsf{pk}}$ with the randomness $r_{\mathcal{A}}$. The reduction provides the adversary with access to its oracles $\mathsf{O}_A = \{\textsc{Reg}, \textsc{Init}, \textsc{Chall}\}$. To answer queries to the $\textsc{Reg}$, oracle the reduction uses the randomness r_R. To answer queries to the $\textsc{Chall}$ oracle, the reduction always uses Fake as described for the current game with the following three changes: First, instead of sampling a new ephemeral KEM key for Bob's pre-key bundle, it uses the next KEM public key from its input list. Second, the reduction replaces lines 35 and 37 (the decapsulation or extraction of the KEM shared secrets) with a call to its own DECAPS oracle of the PA1 game. Third, it uses the randomness r_C where needed. After the adversary terminates, the reduction outputs $(\mathsf{aux}, \vec{\mathsf{pk}}, r, Q, K)$. The PA1-distinguisher $\mathcal{D}'$ then runs the deniability-distinguisher $\mathcal{D}$ on $(\mathsf{aux}, \vec{\mathsf{pk}}, r, Q, K)$ and returns the same guess as the deniability-distinguisher $\mathcal{D}$.

Hence, depending on the secret bit of the PA1 game, the reduction simulates either the previous game or the current game. We include the running time of the $\mathsf{O}_A = \{\text{REG}, \text{INIT}, \text{CHALL}\}$ oracles in $t_{\mathcal{A}}$, allowing for $t_{\mathcal{A}} \approx t_C$.

The PA1 distinguisher $\mathcal{D}'$ executes the deniability distinguisher $\mathcal{D}$, resulting in essentially the same run time $t_{\mathcal{D}} \approx t_{\mathcal{D}^{PA}}$, and directly returns the guess of the deniability distinguisher $\mathcal{D}$. If the deniability distinguisher $\mathcal{D}$ succeeds, then the PA1 distinguisher $\mathcal{D}'$ succeeds as well and can distinguish between the current and the previous game. By the PA1 assumption, the PA1 distinguisher succeeds with an advantage $\leq \epsilon_{\mathcal{D}^{PA}}$, which also limits the advantage of the deniability distinguisher.

Game 2 = Fake *(EKDH assumption).* We replace the honest computations of DH_1, DH_2, DH_3 with the EKDH extractor in lines 26, 27, and 28, finally resulting in Fake as described above. Following Theorem 4.3, we bound the advantage difference introduced by this step by the $(t_{\mathcal{A}^{DH}}, t_{\mathcal{E}^{DH}}, t_{\mathcal{D}^{DH}}, \epsilon_{\mathcal{D}^{DH}})$-EKDH assumption.

If the extractor succeeds (in all three cases), then Fake has computed the session key in the same way as Run, and, hence, indistinguishably. If the extractor fails (in any of the three cases), then the distinguisher succeeds in extracting one of the shared DH secrets with a maximum probability of $\epsilon_{\mathcal{D}^{DH}}$ while running in time $t_{\mathcal{D}^{DH}}$. Hence, the distinguisher can distinguish the real session key from the simulated session key, i.e. the previous game from the current game, with a probability of at most $\epsilon_{\mathcal{D}^{DH}}$.[13] □

For responder deniability we obtain identical guarantees for X3DH and PQXDH. Alice can deny (*responder deniability*) participation in a protocol run against a malicious Bob that honestly generated his keys (*against malicious adversaries with honest keys*, Theorem 4.8), and against a malicious Bob (*against malicious adversaries*, Theorem 4.9). She does so by showing that Bob (1-*out-of*-2, i.e. using Bob's secret keys, Theorem 4.8), or anybody (1-*out-of*-∞, Theorem 4.9) could have produced the transcript and session key himself. This holds even against a big brother distinguisher (*in the big brother model*, Theorem 4.8) and does not rely on eavesdropped info (AuxPrep = ⊥, Theorems 4.8 and 4.9).

THEOREM 4.8. *Both the* X3DH *and* PQXDH *protocols as shown in Figure 3 are* $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* $(\{\text{REGHON}, \text{INIT}, \text{CHALLRESP}\}, \{SK\}, \{SKs\})$*-oracles and auxiliary inputs sampled with* AuxPrep = ⊥, *where* n_p *is the number of parties and* n_{ss} *the number of semi-static keys per party,* $q_{O_A}, q_{O_F}, q_{O_D}$, $t_{\mathcal{A}}, t_{\mathcal{D}}$ *are arbitrary, and* $\epsilon_{\mathcal{D}} = 0$.

THEOREM 4.9 (RESPONDER DENIABILITY AGAINST MALICIOUS ADVERSARIES OF PQXDH). *If the* $(t_{\mathcal{A}^{DH}}, t_{\mathcal{E}^{DH}}, t_{\mathcal{D}^{DH}}, \epsilon_{\mathcal{D}^{DH}})$*-Extended Knowledge of DH (EKDH) Assumption holds and* KDF *is a random oracle, then both the* X3DH *and the* PQXDH *protocols as shown in Figure 3 are* $(n_p, n_{ss}, q_{O_A}, q_{O_F}, q_{O_D}, t_{\mathcal{A}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$*-deniable with respect to* $(\{\text{REG}, \text{INIT}, \text{CHALLRESP}\}, \emptyset, \emptyset)$*-oracles and auxiliary inputs sampled with* AuxPrep = ⊥, *where* n_p *is the number of parties and* n_{ss} *the number of semi-static keys per party,* $q_{O_A}, q_{O_F}, q_{O_D}$ *are arbitrary,* $t_{\mathcal{A}} \approx t_{\mathcal{A}^{DH}}$, $t_{\mathcal{D}} \approx t_{\mathcal{D}^{DH}}$, *and* $\epsilon_{\mathcal{D}} \leq \epsilon_{\mathcal{D}^{DH}}$.

REMARK 4.5 (DENIABILITY AGAINST MALICIOUS ADVERSARIES IN THE BIG BROTHER MODEL). *In Remark 4.2 we have pointed out challenges for showing deniability against malicious adversaries* in the big brother model *for* X3DH. *They directly apply to* PQXDH *as well.*

5 CONCLUSION AND FUTURE WORK

Table 2 summarizes that in comparison to X3DH, PQXDH keeps the deniability guarantees for the responder but not for the initiator: simulating a transcript requires a signature that the distinguisher never sees. It remains an open question if deniability against malicious adversaries holds in the big brother model for both X3DH and PQXDH.

Observe that the AEAD ciphertext influences the assumptions needed to show deniability: Without the AEAD ciphertext, we can show that the DH assumption (and for PQXDH also plaintext awareness of the KEM) is necessary to show deniability against malicious adversaries.[14] When including the AEAD ciphertext, we cannot show the necessity of these assumptions anymore. Also, the KEM used in Signal's implementation, Kyber, does not fulfill plaintext awareness, as pointed out in [38].

One can consider a key awareness assumption (similar to the case of SIGMA [23]) that allows extracting the AEAD key from a malicious Alice. However, on the one hand, the Fake algorithm cannot tell if an extracted AEAD key is the real session key. On the other hand, the distinguisher can tell them apart (since it knows what the adversary does). Observing the adversary's queries to the random oracle does not help either since Fake cannot compute the master secret ms. (Nor could the challenger verify if the adversary queries the random oracle on the actual master secret.) Hence, it appears this proof strategy does not work.

Our analysis treats Bob's long-term DH key and signing key separately, which is not the case in practice. A formal proof for the protocol taking this combined primitive into account is still open.

We see it as an interesting problem to examine if our deniability model is directly applicable to messaging protocols or if any modifications are needed. Cremers and Zhao [18] have extended the deniability notion of [10] for key exchange to messaging protocols. Furthermore, one can extend our deniability model to group chats, see e.g. [31] and the efforts around the MLS standardization process. An extended deniability model for secure messaging can then be used to analyze the deniability guarantees of the whole Signal protocol, i.e. including the Double Ratchet.

[13]Note that the second and third extractor call refer to the same output of the adversary, i.e. they are not independent. Conservatively, we bound with $\epsilon_{\mathcal{D}^{DH}}$ and not $(\epsilon_{\mathcal{D}^{DH}})^3$.

[14]A deniability adversary can use any ciphertext creator to obtain a KEM ciphertext, add a freshly sampled DH public key, and send this as Alice's message to the challenge oracle. We can build a PA extractor by observing the RO queries of the Fake algorithm to learn the KEM shared secret, which must be indistinguishable due to deniability.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive feedback, and Marc Fischlin, Ehren Kret, Trevor Perrin, Olga Sanina, and Rolfe Schmidt for helpful discussions and feedback.

R.F. was supported by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. C.J. was (partially) funded by the Deutsche Forschungsgemeinschaft (DFG) – SFB 1119 – 236615297.

REFERENCES

[1] Karthikean Barghavan, Charlie Jacomme, and Franziskus Kiefer. 2023. Formal Analysis of the PQXDH Protocol. https://github.com/Inria-Prosecco/pqxdh-analysis.

[2] Karthikean Barghavan, Charlie Jacomme, Franziskus Kiefer, and Rolfe Schmidt. 2023. An Analysis of Signal's PQXDH. https://cryspen.com/post/pqxdh/.

[3] Mihir Bellare and Adriana Palacio. 2004. Towards Plaintext-Aware Public-Key Encryption without Random Oracles. In *ASIACRYPT 2004 (LNCS, Vol. 3329)*, Pil Joong Lee (Ed.). Springer, Heidelberg, 48–62. https://doi.org/10.1007/978-3-540-30539-2_4

[4] Mihir Bellare and Phillip Rogaway. 1995. Optimal Asymmetric Encryption. In *EUROCRYPT'94 (LNCS, Vol. 950)*, Alfredo De Santis (Ed.). Springer, Heidelberg, 92–111. https://doi.org/10.1007/BFb0053428

[5] Nikita Borisov, Ian Goldberg, and Eric A. Brewer. 2004. Off-the-record communication, or, why not to use PGP. In *Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society, WPES 2004, Washington, DC, USA, October 28, 2004*, Vijay Atluri, Paul F. Syverson, and Sabrina De Capitani di Vimercati (Eds.). ACM, 77–84. https://doi.org/10.1145/1029179.1029200

[6] Colin Boyd, Wenbo Mao, and Kenneth G. Paterson. 2003. Deniable Authenticated Key Establishment for Internet Protocols. In *Security Protocols Workshop*. 255–271.

[7] Colin Boyd, Wenbo Mao, and Kenneth G. Paterson. 2004. Key Agreement Using Statically Keyed Authenticators. In *ACNS 04 (LNCS, Vol. 3089)*, Markus Jakobsson, Moti Yung, and Jianying Zhou (Eds.). Springer, Heidelberg, 248–262. https://doi.org/10.1007/978-3-540-24852-1_18

[8] Colin Boyd, Anish Mathuria, and Douglas Stebila. 2020. *Protocols for Authentication and Key Establishment, Second Edition*. Springer. https://doi.org/10.1007/978-3-662-58146-9

[9] Jacqueline Brendel, Rune Fiedler, Felix Günther, Christian Janson, and Douglas Stebila. 2021. Post-quantum Asynchronous Deniable Key Exchange and the Signal Handshake. Cryptology ePrint Archive, Report 2021/769. https://eprint.iacr.org/2021/769.

[10] Jacqueline Brendel, Rune Fiedler, Felix Günther, Christian Janson, and Douglas Stebila. 2022. Post-quantum Asynchronous Deniable Key Exchange and the Signal Handshake. In *PKC 2022, Part II (LNCS, Vol. 13178)*, Goichiro Hanaoka, Junji Shikata, and Yohei Watanabe (Eds.). Springer, Heidelberg, 3–34. https://doi.org/10.1007/978-3-030-97131-1_1

[11] Jacqueline Brendel, Marc Fischlin, Felix Günther, Christian Janson, and Douglas Stebila. 2020. Towards Post-Quantum Security for Signal's X3DH Handshake. In *SAC 2020 (LNCS, Vol. 12804)*, Orr Dunkelman, Michael J. Jacobson Jr., and Colin O'Flynn (Eds.). Springer, Heidelberg, 404–430. https://doi.org/10.1007/978-3-030-81652-0_16

[12] Ran Canetti and Hugo Krawczyk. 2002. Security Analysis of IKE's Signature-based Key-Exchange Protocol. In *CRYPTO 2002 (LNCS, Vol. 2442)*, Moti Yung (Ed.). Springer, Heidelberg, 143–161. https://doi.org/10.1007/3-540-45708-9_10 https://eprint.iacr.org/2002/120/.

[13] Wouter Castryck and Thomas Decru. 2023. An Efficient Key Recovery Attack on SIDH. In *EUROCRYPT 2023, Part V (LNCS, Vol. 14008)*, Carmit Hazay and Martijn Stam (Eds.). Springer, Heidelberg, 423–447. https://doi.org/10.1007/978-3-031-30589-4_15

[14] Katriel Cohn-Gordon, Cas J. F. Cremers, Benjamin Dowling, Luke Garratt, and Douglas Stebila. 2017. A Formal Security Analysis of the Signal Messaging Protocol. In *IEEE European Symposium on Security and Privacy, EuroS&P 2017*. 451–466. https://doi.org/10.1109/EuroSP.2017.27

[15] Daniel Collins, Simone Colombo, and Loïs Huguenin-Dumittan. 2023. Real World Deniability in Messaging. Cryptology ePrint Archive, Paper 2023/403. https://eprint.iacr.org/2023/403

[16] Daniel Collins, Loïs Huguenin-Dumittan, Ngoc Khanh Nguyen, Nicolas Rolin, and Serge Vaudenay. 2024. K-Waay: Fast and Deniable Post-Quantum X3DH without Ring Signatures. In *33rd USENIX Security Symposium (USENIX Security 24)*. USENIX Association, Philadelphia, PA. https://www.usenix.org/conference/usenixsecurity24/presentation/collins

[17] Cas Cremers and Michele Feltz. 2011. One-round Strongly Secure Key Exchange with Perfect Forward Secrecy and Deniability. Cryptology ePrint Archive, Report 2011/300. https://eprint.iacr.org/2011/300.

[18] Cas Cremers and Mang Zhao. 2024. Secure Messaging with Strong Compromise Resilience, Temporal Privacy, and Immediate Decryption. In *IEEE Symposium on Security and Privacy*.

[19] Özgür Dagdelen, Marc Fischlin, Tommaso Gagliardoni, Giorgia Azzurra Marson, Arno Mittelbach, and Cristina Onete. 2013. A Cryptographic Analysis of OPACITY - (Extended Abstract). In *ESORICS 2013 (LNCS, Vol. 8134)*, Jason Crampton, Sushil Jajodia, and Keith Mayes (Eds.). Springer, Heidelberg, 345–362. https://doi.org/10.1007/978-3-642-40203-6_20

[20] Alexander W. Dent. 2006. The Cramer-Shoup Encryption Scheme Is Plaintext Aware in the Standard Model. In *EUROCRYPT 2006 (LNCS, Vol. 4004)*, Serge Vaudenay (Ed.). Springer, Heidelberg, 289–307. https://doi.org/10.1007/11761679_18

[21] Mario Di Raimondo and Rosario Gennaro. 2005. New Approaches for Deniable Authentication. In *ACM CCS 2005*, Vijayalakshmi Atluri, Catherine Meadows, and Ari Juels (Eds.). ACM Press, 112–121. https://doi.org/10.1145/1102120.1102137

[22] Mario Di Raimondo, Rosario Gennaro, and Hugo Krawczyk. 2005. Secure off-the-record messaging. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, WPES 2005, Alexandria, VA, USA, November 7, 2005*, Vijay Atluri, Sabrina De Capitani di Vimercati, and Roger Dingledine (Eds.). ACM, 81–89. https://doi.org/10.1145/1102199.1102216

[23] Mario Di Raimondo, Rosario Gennaro, and Hugo Krawczyk. 2006. Deniable authentication and key exchange. In *ACM CCS 2006*, Ari Juels, Rebecca N. Wright, and Sabrina De Capitani di Vimercati (Eds.). ACM Press, 400–409. https://doi.org/10.1145/1180405.1180454

[24] Samuel Dobson and Steven D. Galbraith. 2022. Post-Quantum Signal Key Agreement from SIDH. In *Post-Quantum Cryptography - 13th International Workshop, PQCrypto 2022, Virtual Event, September 28-30, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13512)*, Jung Hee Cheon and Thomas Johansson (Eds.). Springer, 422–450. https://doi.org/10.1007/978-3-031-17234-2_20

[25] Yevgeniy Dodis, Jonathan Katz, Adam Smith, and Shabsi Walfish. 2009. Composability and On-Line Deniability of Authentication. In *TCC 2009 (LNCS, Vol. 5444)*, Omer Reingold (Ed.). Springer, Heidelberg, 146–162. https://doi.org/10.1007/978-3-642-00457-5_10

[26] Cynthia Dwork, Moni Naor, and Amit Sahai. 1998. Concurrent Zero-Knowledge. In *30th ACM STOC*. ACM Press, 409–418. https://doi.org/10.1145/276698.276853

[27] Rune Fiedler and Felix Günther. 2024. Security Analysis of Signal's PQXDH Handshake. Cryptology ePrint Archive, Paper 2024/702. https://eprint.iacr.org/2024/702.

[28] Dan Harkins, Charlie Kaufman, Tero Kivinen, Stephen Kent, and Radia Perlman. 2002. *Design Rationale for IKEv2*. Internet-Draft draft-ietf-ipsec-ikev2-rationale-00.txt. IETF Secretariat. https://www.ietf.org/proceedings/54/I-D/draft-ietf-ipsec-ikev2-rationale-00.txt

[29] Lein Harn, Chia-Yin Lee, Changlu Lin, and Chin-Chen Chang. 2011. Fully Deniable Message Authentication Protocols Preserving Confidentiality. *Comput. J.* 54, 10 (2011), 1688–1699. https://doi.org/10.1093/comjnl/bxr081

[30] Keitaro Hashimoto, Shuichi Katsumata, Kris Kwiatkowski, and Thomas Prest. 2022. An Efficient and Generic Construction for Signal's Handshake (X3DH): Post-quantum, State Leakage Secure, and Deniable. *Journal of Cryptology* 35, 3 (July 2022), 17. https://doi.org/10.1007/s00145-022-09427-1

[31] Andreas Hülsing and Fiona Johanna Weber. 2021. Epochal Signatures for Deniable Group Chats. In *2021 IEEE Symposium on Security and Privacy*. IEEE Computer Society Press, 1677–1695. https://doi.org/10.1109/SP40001.2021.00058

[32] Shaoquan Jiang. 2014. Timed encryption with application to deniable key exchange. *Theor. Comput. Sci.* 560 (2014), 172–189. https://doi.org/10.1016/J.TCS.2014.02.005

[33] Shaoquan Jiang, Yeow Meng Chee, San Ling, Huaxiong Wang, and Chaoping Xing. 2022. A new framework for deniable secure key exchange. *Inf. Comput.* 285, Part (2022), 104866. https://doi.org/10.1016/J.IC.2022.104866

[34] Shaoquan Jiang and Reihaneh Safavi-Naini. 2008. An Efficient Deniable Key Exchange Protocol (Extended Abstract). In *FC 2008 (LNCS, Vol. 5143)*, Gene Tsudik (Ed.). Springer, Heidelberg, 47–52.

[35] Shaoquan Jiang and Huaxiong Wang. 2010. Plaintext-Awareness of Hybrid Encryption. In *CT-RSA 2010 (LNCS, Vol. 5985)*, Josef Pieprzyk (Ed.). Springer, Heidelberg, 57–72. https://doi.org/10.1007/978-3-642-11925-5_5

[36] Jonathan Katz. 2003. Efficient and Non-malleable Proofs of Plaintext Knowledge and Applications. In *EUROCRYPT 2003 (LNCS, Vol. 2656)*, Eli Biham (Ed.). Springer, Heidelberg, 211–228. https://doi.org/10.1007/3-540-39200-9_13

[37] Hugo Krawczyk. 1996. SKEME: a versatile secure key exchange mechanism for Internet. In *NDSS'96*, James T. Ellis, B. Clifford Neuman, and David M. Balenson (Eds.). IEEE Computer Society, 114–127. https://doi.org/10.1109/NDSS.1996.492418

[38] Ehren Kret and Rolfe Schmidt. September 2023. The PQXDH key agreement protocol. https://signal.org/docs/specifications/pqxdh/.

[39] Meng-Hui Lim, Sanggon Lee, Youngho Park, and Sangjae Moon. 2007. Secure Deniable Authenticated Key Establishment for Internet Protocols. Cryptology

ePrint Archive, Report 2007/163. https://eprint.iacr.org/2007/163.
[40] Luciano Maino, Chloe Martindale, Lorenz Panny, Giacomo Pope, and Benjamin Wesolowski. 2023. A Direct Key Recovery Attack on SIDH. In *EUROCRYPT 2023, Part V (LNCS, Vol. 14008)*, Carmit Hazay and Martijn Stam (Eds.). Springer, Heidelberg, 448–471. https://doi.org/10.1007/978-3-031-30589-4_16
[41] Wenbo Mao and Kenneth G. Paterson. 2002. On the Plausible Deniability Feature of Internet Protocols. (2002). unpublished, https://web.archive.org/web/20220818192033/http://www.isg.rhul.ac.uk/~kp/IKE.ps.
[42] Moxie Marlinspike and Trevor Perrin. November 2016. The double ratchet algorithm. https://www.signal.org/docs/specifications/doubleratchet/.
[43] Moxie Marlinspike and Trevor Perrin. November 2016. The X3DH key agreement protocol. https://signal.org/docs/specifications/x3dh/.
[44] Moni Naor. 2002. Deniable Ring Authentication. In *CRYPTO 2002 (LNCS, Vol. 2442)*, Moti Yung (Ed.). Springer, Heidelberg, 481–498. https://doi.org/10.1007/3-540-45708-9_31
[45] OTR team. 2020. *OTR version 4*. https://github.com/otrv4/otrv4/blob/master/otrv4.md.
[46] Rafael Pass. 2003. On Deniability in the Common Reference String and Random Oracle Model. In *CRYPTO 2003 (LNCS, Vol. 2729)*, Dan Boneh (Ed.). Springer, Heidelberg, 316–337. https://doi.org/10.1007/978-3-540-45146-4_19
[47] Trevor Perrin. October 2016. The XEdDSA and VXEdDSA Signature Schemes. https://signal.org/docs/specifications/xeddsa/.
[48] Nathan Reitinger, Nathan Malkin, Omer Akgul, Michelle L. Mazurek, and Ian Miers. 2023. Is Cryptographic Deniability Sufficient*f* Non-Expert Perceptions of Deniability in Secure Messaging. In *2023 IEEE Symposium on Security and Privacy*. IEEE Computer Society Press, 274–292. https://doi.org/10.1109/SP46215.2023.10179361
[49] Damien Robert. 2023. Breaking SIDH in Polynomial Time. In *EUROCRYPT 2023, Part V (LNCS, Vol. 14008)*, Carmit Hazay and Martijn Stam (Eds.). Springer, Heidelberg, 472–503. https://doi.org/10.1007/978-3-031-30589-4_17
[50] Phillip Rogaway. 2002. Authenticated-Encryption With Associated-Data. In *ACM CCS 2002*, Vijayalakshmi Atluri (Ed.). ACM Press, 98–107. https://doi.org/10.1145/586110.586125
[51] Peter Schwabe, Roberto Avanzi, Joppe Bos, Léo Ducas, Eike Kiltz, Tancrède Lepoint, Vadim Lyubashevsky, John M. Schanck, Gregor Seiler, Damien Stehlé, and Jintai Ding. 2022. *CRYSTALS-KYBER.* Technical Report. National Institute of Standards and Technology. available at https://csrc.nist.gov/Projects/post-quantum-cryptography/selected-algorithms-2022.
[52] Nik Unger. 2021. *End-to-End Encrypted Group Messaging with Insider Security.* Ph.D. Dissertation. University of Waterloo, Ontario, Canada. https://hdl.handle.net/10012/17196.
[53] Nik Unger, Sergej Dechand, Joseph Bonneau, Sascha Fahl, Henning Perl, Ian Goldberg, and Matthew Smith. 2015. SoK: Secure Messaging. In *2015 IEEE Symposium on Security and Privacy*. IEEE Computer Society Press, 232–249. https://doi.org/10.1109/SP.2015.22
[54] Nik Unger and Ian Goldberg. 2015. Deniable Key Exchanges for Secure Messaging. In *ACM CCS 2015*, Indrajit Ray, Ninghui Li, and Christopher Kruegel (Eds.). ACM Press, 1211–1223. https://doi.org/10.1145/2810103.2813616
[55] Nik Unger and Ian Goldberg. 2018. Improved Strongly Deniable Authenticated Key Exchanges for Secure Messaging. *PoPETs* 2018, 1 (Jan. 2018), 21–66. https://doi.org/10.1515/popets-2018-0003
[56] Nihal Vatandas, Rosario Gennaro, Bertrand Ithurburn, and Hugo Krawczyk. 2020. On the Cryptographic Deniability of the Signal Protocol. In *ACNS 20, Part II (LNCS, Vol. 12147)*, Mauro Conti, Jianying Zhou, Emiliano Casalicchio, and Angelo Spognardi (Eds.). Springer, Heidelberg, 188–209. https://doi.org/10.1007/978-3-030-57878-7_10
[57] Tarun Kumar Yadav, Devashish Gosain, and Kent E. Seamons. 2023. Cryptographic Deniability: A Multi-perspective Study of User Perceptions and Expectations. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, Joseph A. Calandrino and Carmela Troncoso (Eds.). USENIX Association, 3637–3654. https://www.usenix.org/conference/usenixsecurity23/presentation/yadav
[58] Andrew Chi-Chih Yao and Yunlei Zhao. 2010. Deniable Internet Key Exchange. In *ACNS 10 (LNCS, Vol. 6123)*, Jianying Zhou and Moti Yung (Eds.). Springer, Heidelberg, 329–348. https://doi.org/10.1007/978-3-642-13708-2_20
[59] Andrew Chi-Chih Yao and Yunlei Zhao. 2013. OAKE: a new family of implicitly authenticated Diffie-Hellman protocols. In *ACM CCS 2013*, Ahmad-Reza Sadeghi, Virgil D. Gligor, and Moti Yung (Eds.). ACM Press, 1113–1128. https://doi.org/10.1145/2508859.2516695

A COMPARING PRIOR DEFINITIONS TO OUR MODEL

This section discusses in more detail how our model captures the idea of prior definitions from the literature (sketched by Table 3). Keep in mind that whenever we write $\textsc{Chall}$ we refer to the challenge oracles for both roles (i.e. $\textsc{ChallInit}, \textsc{ChallResp}$), and similarly for $\textsc{ChallHon}$ and $\textsc{ChallHonInit}, \textsc{ChallHonResp}$.

The following deniability models are not concerned with user messages. Though, we expect that they can be adapted similar to our model to handle key exchange protocols with user messages.

A.1 Concurrent and partial deniability [23]

Deniability in the context of key exchange was introduced by Di Raimondo, Gennaro and Krawczyk [23], namely concurrent deniability and partial deniability. Let us first focus on *concurrent deniability* [23, Definition 2]. Both definitions are simulation-based definitions, i.e. a distinguisher has to decide if it is presented with values sampled from the real distribution or the simulated distribution. In either case, key pairs for all honest users and auxiliary info (which may depend on the keys) are sampled first. On a technical note, the probability of the distinguisher succeeding is taken over the randomness in generating the keys (and consequentially the sampling of the auxiliary info) and the random coins of the adversary. Note that the distinguisher gets access to the auxiliary info as well.

In the real case, the adversary may then interact with oracles representing the honest users in an arbitrary manner; the output consists of the previously sampled public keys and auxiliary info, the randomness that the adversary used, the transcripts between the adversary and all oracles, and the session keys of all completed sessions. The astute observer notices the similarity to our game-based model with $b = 0$ for $\mathrm{O}_A = \{\textsc{Reg}, \textsc{Init}, \textsc{Chall}\}$ (with syntactic changes). In the simulated case, a simulator (who depends on the adversary) is given the same input as the adversary (i.e. the public keys, the auxiliary info, and the randomness that the adversary used) and needs to produce indistinguishable output. Note that the distinguisher receives the auxiliary info as well.

The proof technique for this definition suggested by [23] and used thenceforth [17, 30, 33, 56] is that the simulator executes the adversary and answers in place of the oracle. Subsequent work [32, 34] follows this approach and directly tasks the simulator to answer in place of the oracles. The advantage of this technique is that the final transcript then contains identically distributed messages by the adversary. Our game-based definition covers this case with $b = 1$ by using the Fake algorithm for simulation. To be more precise, the Fake algorithm of our game-based definition depends only on the currently challenged session and does not have a shared state over all invocations. The authors believe that this is also intended by [23], who focus on concurrent sessions, where the simulation needs to happen regardless of other sessions. However, they need the simulator to have an overall state to manage the queries of the adversary.

Hence, we capture the idea of their concurrent deniability with $\mathrm{O}_A = \{\textsc{Reg}, \textsc{Init}, \textsc{Chall}\}, \mathrm{O}_F = \emptyset, \mathrm{O}_D = \{\textsc{aux}\}$, and arbitrary $\mathsf{AuxPrep}$.

Their *partial deniability* [23, Definition 3] allows the simulator to access an oracle acting as the challenged user partnered with another user. So we capture the idea of their partial deniability with $\mathrm{O}_A = \{\textsc{Reg}, \textsc{Init}, \textsc{Chall}\}, \mathrm{O}_F = \{\textsc{User}_{x,y}\}, \mathrm{O}_D = \{\textsc{aux}\}$, arbitrary $\mathsf{AuxPrep}$, and $q_U = 1$, i.e. per session for which the Fake algorithm

Type of Deniability	O_A	O_F	O_D	Comment
concurrent deniability [23]	Reg, Init, Chall	$\emptyset$	aux	
partial deniability [23]	Reg, Init, Chall	$\text{User}_{x,y}$	aux	$q_U = 1$
peer deniability [17]	Reg, RegHon, Init, Chall	$\text{User}_{x,y}$, SK	$\emptyset$	AuxPrep = $\perp$, session keys not deniable
peer-and-time deniability [17]	Reg, RegHon, Init, Chall	SK	$\emptyset$	AuxPrep has transcripts, session keys not deniable
deniability (against malicious adv.) [30]	Reg, Init, Chall	$\emptyset$	$\emptyset$	AuxPrep = $\perp$
deniability (against semi-honest adv.) [30]	RegHon, ChallHon	$\emptyset$	$\emptyset$	AuxPrep = $\perp$
outsider deniability [19]	ChallPassive	$\emptyset$	$\emptyset$	AuxPrep has valid transcripts
deniability [9, 10]	ChallPassive	SK	SKs	AuxPrep = $\perp$
HP-deniability [59]	ChallPassive	$\emptyset$	$\emptyset$	AuxPrep = $\perp$
$DENY^{\text{false}}$ and $DENY^{\text{true}}$ [16]	ChallPassive	SK	SKs	AuxPrep = $\perp$
deniability [34]	Reg, Init, Chall	$\emptyset$	aux	AuxPrep has secret keys
deniability [32]	Reg, Init, Chall	$\emptyset$	aux	AuxPrep has secret keysandvalid transcripts
deniability [33]	Reg, Init, Chall	$\emptyset$	aux	AuxPrep has secret keysandvalid transcripts

Table 3: Relation between our model and prior work: How to achieve deniability notions similar to those in the literature. Challenge oracles not suffixed by a role refer to the oracles for both roles. Deniability [9, 10], outsider deniability [19], $DENY^{\text{false}}$ and $DENY^{\text{true}}$ [16] are game-based definitions; the others are simulation-based.

is queried it can start one session with the target user (who is then partnered with a new, honest user).

A.2 Peer deniability and peer-and-time deniability [17]

Peer deniability and peer-and-time deniability [17] are simulation-based notions: The former is based on partial deniability [23] and the latter is a strengthening of peer deniability. Neither definition uses auxiliary info.

A.2.1 Peer deniability. Starting from partial deniability [23], peer deniability [17, Definition 9] is conceptually close with a few important differences:

- Peer deniability does not guarantee deniability of session keys. Hence, in line 7 of Figure 2 the distinguisher does not get access to the session keys computed by the Chall oracle.
- The simulator can access polynomially many sessions to create messages, i.e. q_U is not fixed to 1.
- They give the attacker access to a corruption oracle, which allows the attacker *and the simulator* to learn the secret key(s) of the corrupted parties. We model the corruption oracle with the RegHon and SK oracles: The former allows the adversary to learn the secret key of a honestly generated key pair, and the latter for the Fake algorithm.[15] However, we retain the Reg oracle from partial deniability [23]. In consequence, the Fake algorithm cannot rely on the SK oracle, just like the simulator cannot rely on learning the peer's secret key from the set of corrupted parties (maybe the adversary does not corrupt the party in question).

Hence, we capture the idea of their definition by setting $O_A = \{\text{Init}, \text{Reg}, \text{RegHon}, \text{Chall}\}$, $O_F = \{\text{User}_{x,y}, \text{SK}\}$, $O_D = \emptyset$, and AuxPrep $= \perp$. Additionally, we remove K as argument to $\mathcal{D}$ in line 7 of Figure 2.

A.2.2 Peer-and-time deniability. Peer-and-time deniability [17, Definition 10] strengthens peer deniability (above) by restricting the simulator (i.e. the Fake oracle): The simulator has to finish interacting with the $\text{User}_{x,y}$ oracle before interacting with the adversary. This ensures that the victim doesn't produce incriminating messages dependent on the adversary's input.

For our game-based approach we split up the simulator into many calls to Fake, each producing a single message. Hence, we cannot directly adapt the notion of Fake first interacting with a user oracle as preparation, and replying to the adversary only afterwards. We note, however, that peer-and-time deniability does not use auxiliary info, allowing us to repurpose the auxiliary info in our game-based definition: Originally, the simulator uses some strategy to query its oracle in an arbitrary manner. We encode this strategy into the AuxPrep algorithm and give aux only to Fake (and not to $\mathcal{A}$).

Both approaches (the simulator having access to an oracle before interacting with the adversary and our tailored sampling of auxiliary info) provide auxiliary info to Fake that the distinguisher is not aware of. Therefore, the Fake algorithm can use that info without the distinguisher noticing.

Hence, we capture the idea of their definition by setting $O_A = \{\text{Init}, \text{Reg}, \text{RegHon}, \text{Chall}\}$, $O_F = \{\text{SK}\}$, $O_D = \emptyset$, and AuxPrep follows the simulator's strategy for querying $\text{User}_{x,y}$. Additionally, we remove K as argument to $\mathcal{D}$ in line 7 of Figure 2 and aux as argument to $\mathcal{A}$ in line 6 in Figure 2.

A.3 Deniability against malicious and semi-honest adversaries [30]

Deniability against malicious and semi-honest adversaries [30, Definition 7.1] are two simulation-based notions based on [23]. Neither case allows any auxiliary information.

[15]Strictly speaking, they give all corrupted secret keys to the simulator, while we give only the peer's secret key to the Fake algorithm. However, the authors do not see how a third party's secret key would be of help.

To capture the idea of their definition against malicious adversaries, we follow the same approach as for [23] and set $\mathrm{O}_{\mathcal{A}} = \{\textsc{Reg}, \textsc{Init}, \textsc{Chall}\}, \mathrm{O}_F = \emptyset, \mathrm{O}_D = \emptyset$, and AuxPrep $= \perp$. To capture the idea of their definition against semi-honest adversaries, we limit the adversary to accessing the oracles which enforce adherence to the protocol flow. Hence, we set $\mathrm{O}_{\mathcal{A}} = \{\textsc{RegHon}, \textsc{ChallHon}\}$, $\mathrm{O}_F = \emptyset, \mathrm{O}_D = \emptyset$, and AuxPrep $= \perp$.

A.4 Outsider deniability [19]

Outsider deniability [19, Definition 2.5] is a game based-notion. Their notion gives the adversary access to an execute oracle (which yields a transcript) and a test oracle, which - depending on the secret bit - yields either a real transcript and session key or simulated values. Unlike our Fake algorithm, their simulator produces the complete transcript at once. Their notion is concerned with a passive adversary and a transcript should be simulatable with public data only. They do not allow auxiliary info, nor does the distinguisher get the adversary's randomness.

Hence, to capture the idea of their definition we set $\mathrm{O}_{\mathcal{A}} = \{\textsc{ChallPassive}\}, \mathrm{O}_F = \emptyset, \mathrm{O}_D = \emptyset$, and AuxPrep contains valid transcripts (substituting for the execute oracle). Additionally, we remove r as argument to $\mathcal{D}$ in line 7 of Figure 2.

A.5 Deniability [9, 10]

Deniability [9, Definition 11] is a game-based notion of 1-out-of-2 deniability against semi-honest adversaries in the big brother model. They tailor the deniability notion to their use case of a replacement for Signal's initial handshake: They argue that Bob's pre-key bundle does not need to be deniable since it is not bound to any peer or any session. Hence, it suffices for Alice's message to be simulatable by Bob to achieve 1-out-of-2 deniability.

Note that they consider Bob, who prepares the pre-key bundle, the responder, while we consider Bob the initiator.

Furthermore, they merge the adversary and the distinguisher into one algorithm $\mathcal{A}$. Syntactically, this has the consequence that the adversary-distinguisher already learns all secret keys while having access to the challenge oracle. Since the adversary-distinguisher is semi-honest, we are not aware how the knowledge of secret keys could help the adversary-distinguisher.

While they give a Fake algorithm, their Fake algorithm syntactically differs from ours: It directly produces a complete transcript and session key. Semantically this is equivalent for the use within our ChallPassive oracle. They do not use auxiliary info.

All in all, to capture the idea of their definition we set $\mathrm{O}_{\mathcal{A}} = \{\textsc{ChallPassive}\}, \mathrm{O}_F = \{\mathrm{SK}\}, \mathrm{O}_D = \{\mathrm{SKs}\}$, and AuxPrep $= \perp$, where the SK oracle only responds to queries for the secret keys of the initiator Bob.

A.6 HP-deniability [59]

HP-deniability (short for honest-player-deniability) [59, Definition 6.1] is a simulation-based definition that targets passive adversaries and exposes "pre-computed and stored session-states" to the distinguisher. In [59] the authors stress that this ephemeral state is not necessarily equivalent to the random coins used for the session. We observe that some protocols, e.g. X3DH and PQXDH, instruct parties to delete their ephemeral secrets after use. Hence, it is not generally obvious which information should be included in this ephemeral session state, but appears to be protocol-specific.

All in all, to capture the idea of their definition we set $\mathrm{O}_{\mathcal{A}} = \{\textsc{ChallPassive}\}, \mathrm{O}_F = \emptyset, \mathrm{O}_D = \emptyset$, and AuxPrep $= \perp$. Additionally, we need an extra game variable to save the ephemeral session state in: It is initialized in line 1 of Figure 2, populated in line 20 at the end of the ChallPassive oracle in Figure 6, and is given to the distinguisher in line 7 of Figure 2.

A.7 $DENY^{\text{false}}$ and $DENY^{\text{true}}$ [16]

The $DENY^{\text{false}}$ notion adapts the idea of deniability [9, 10] to the formalization from [16]. Similarly, it is another game-based notion of 1-out-of-2 deniability against semi-honest adversaries in the big brother model. The $DENY^{\text{true}}$ notion extends $DENY^{\text{false}}$ by additionally giving the adversary-distinguisher access to the responder's state. They remark that the $DENY^{\text{false}}$ notion intuitively models the responder trying to frame the initiator, and the $DENY^{\text{true}}$ notion adds that the responder, who frames the initiator, cooperates with the judge by handing over his or her ephemeral state after the protocol run. As for HP-deniability [59], we remark that some protocols, e.g. X3DH and PQXDH, require users to delete ephemeral data after completing the protocol run. Hence, it appears that the ephemeral session state needs to be defined and included on a per-protocol basis. They do not use auxiliary info.

All in all, to capture the idea of their definition we set $\mathrm{O}_{\mathcal{A}} = \{\textsc{ChallPassive}\}, \mathrm{O}_F = \{\mathrm{SK}\}, \mathrm{O}_D = \{\mathrm{SKs}\}$, and AuxPrep $= \perp$, where the SK oracle only responds to queries for the secret keys of the initiator Bob.

A.8 Deniability [34]

Deniability [34] is a simulation-based definition that allows adversarial corruptions. Note that adversary, simulator, and distinguisher all learn the corrupted secret keys. Hence, we model this non-adaptively by encoding the adversary's corruption strategy into AuxPrep, which then yields the corresponding secret keys to all three algorithms $\mathcal{A}$, Fake, and $\mathcal{D}$. While their definition leaves it implicit, we assume they allow the adversary to create keys maliciously.

To capture the idea of their definition we set $\mathrm{O}_{\mathcal{A}} = \{\textsc{Reg}, \textsc{Init}, \textsc{ChallInit}, \textsc{ChallResp}\}, \mathrm{O}_F = \emptyset, \mathrm{O}_D = \{\textsc{aux}\}$, and AuxPrep yields secret keys according to a given corruption strategy.

A.9 Deniability [32]

Deniability [32] is similar to the notion of [34] but additionally allows access to valid transcripts for all three algorithms $\mathcal{A}$, Fake, and $\mathcal{D}$.

To capture the idea of their definition we set $\mathrm{O}_{\mathcal{A}} = \{\textsc{Reg}, \textsc{Init}, \textsc{ChallInit}, \textsc{ChallResp}\}, \mathrm{O}_F = \emptyset, \mathrm{O}_D = \{\textsc{aux}\}$, and AuxPrep yields secret keys according to a given corruption strategy and valid protocol transcripts.

A.10 Deniability [33]

Deniability [33], follows the ideas of [32]. Additionally, they give the adversary access to a Reveal oracle, which yields the session key. We do not need this extra oracle since our model provides the session key to the adversary and distinguisher directly.

We can capture the idea of their definition as for [32].

B DEFERRED PRELIMINARIES

We give the deferred preliminaries.

B.1 Signatures

We briefly review the syntax for digital signature schemes.

Definition B.1 (Signature schemes). A *digital signature scheme* is a triple of algorithms SIG = (KGen, Sig, Vf) with associated message space $\mathcal{M}_{\mathsf{SIG}}$, defined as follows:

- $\mathsf{KGen}() \xrightarrow{\$} (\mathsf{pk}, \mathsf{sk})$: This probabilistic algorithm returns a key pair (pk, sk)
- $\mathsf{Sign}(\mathsf{sk}, m) \xrightarrow{\$} \sigma$: On input a secret key sk and a message $m \in \mathcal{M}_{\mathsf{SIG}}$, this probabilistic algorithm returns a signature σ;
- $\mathsf{Vf}(\mathsf{pk}, m, \sigma) \to d$: On input a public verification key pk, a message m, and a candidate signature σ, this deterministic algorithm returns a bit $d \in \{0, 1\}$. If $d = 1$ we say that the signature is valid, otherwise not.

We say that a digital signature scheme SIG is *correct* if, for every $(\mathsf{pk}, \mathsf{sk}) \xleftarrow{\$} \mathsf{KGen}()$, every $m \in \mathcal{M}_{\mathsf{SIG}}$, and random $\sigma \xleftarrow{\$} \mathsf{Sign}(\mathsf{sk}, m)$, it holds that $\Pr[\mathsf{Vf}(\mathsf{pk}, m, \sigma) = 1] = 1$.

We consider the security notion unforgeability under chosen message attacks for digital signature schemes.

Definition B.2 (Unforgeability for digital signature Schemes). A digital signature scheme SIG is $(q_{\mathsf{SIG}}, t_{\mathsf{SIG}}, \epsilon_{\mathsf{SIG}})$-unforgeable if, for every adversary $\mathcal{A}$ running in time t_{SIG} and making at most q_{SIG} queries to the SIGN oracle, it holds that $\Pr[\mathcal{G}^{uf}_{\mathsf{SIG}}(\mathcal{A}) = 1] \leq \epsilon_{\mathsf{SIG}}$, where $\mathcal{G}^{uf}_{\mathsf{SIG}}(\mathcal{A})$ is defined as:

```
$\mathcal{G}^{uf}_{\mathsf{SIG}}(\mathcal{A})$:
1 $Q \leftarrow \emptyset$
2 $(\mathsf{pk}, \mathsf{sk}) \xleftarrow{\$} \mathsf{SIG.KGen}()$
3 $(m^*, \sigma^*) \xleftarrow{\$} \mathcal{A}^{\textsc{Sign}}(\mathsf{pk})$
4 return $\mathsf{Vf}(\mathsf{pk}, m^*, \sigma^*) \wedge m^* \notin Q$

$\textsc{Sign}(m)$:
5 $\sigma \xleftarrow{\$} \mathsf{Sign}(\mathsf{sk}, m)$
6 $Q \leftarrow Q \cup \{m\}$
7 return $\sigma$
```

We briefly review the syntax for the Diffie–Hellman key exchange.

Definition B.3 (Diffie–Hellman Key Exchange). A *Diffie–Hellman Key Exchange (DH) scheme* is a tuple of algorithms (KGen, DH), defined as follows:

- $\mathsf{KGen}() \xrightarrow{\$} (\mathsf{pk}, \mathsf{sk})$: This probabilistic algorithm returns a key pair (pk, sk)
- $\mathsf{DH}(\mathsf{pk}_A, \mathsf{sk}_B) \to \mathsf{DH}_{AB}$: On input a public key pk_A and a secret key sk_B, this deterministic algorithm returns the shared DH secret DH_{AB}.

We say that a DH key exchange DH is *correct* if, for every $(\mathsf{pk}_A, \mathsf{sk}_A), (\mathsf{pk}_B, \mathsf{sk}_B) \xleftarrow{\$} \mathsf{KGen}()$, it holds that $\Pr[\mathsf{DH}(\mathsf{pk}_A, \mathsf{sk}_B) = \mathsf{DH}(\mathsf{pk}_B, \mathsf{sk}_A)] = 1$. For notational convenience we allow the arguments to be in arbitrary order. Hence, $\mathsf{DH}(\mathsf{pk}_A, \mathsf{sk}_B) = \mathsf{DH}(\mathsf{sk}_A, \mathsf{pk}_B)$.

The following assumption differs slightly from the EKDH assumption in Definition 2.1: Here, $\mathcal{A}$ is first given two DH public keys and outputs a third DH public key. Then, the extractor tries to extract the two DH shared secrets between the output public key and each of the input public keys.

Definition B.4 (K2DH Assumption [56]). Let G be a cyclic group with generator g and AuxPrep a sampler for auxiliary inputs. Let $\mathcal{A}$ be any algorithm running in time $t_{\mathcal{A}}$ which runs on input (U, W, aux) where $U, W \in G$, and $\mathsf{aux} \xleftarrow{\$} \mathsf{AuxPrep}$, and outputs $Z \in G$; we denote with $Z = \mathcal{A}(U, W, \mathsf{aux}, r)$ the output of running $\mathcal{A}$ on input U, W, aux with coins r.

We say that the $(t_{\mathcal{A}}, t_{\mathcal{E}}, t_{\mathcal{D}}, \epsilon_{\mathcal{D}})$-Knowledge of 2DH (K2DH) Assumption holds over group G and for auxiliary info AuxPrep, if for every such $\mathcal{A}$, there exists a companion algorithm $\mathcal{E}^{DH}_{\mathcal{A}}$ (called the extractor for $\mathcal{A}$) such that: $\mathcal{E}^{DH}_{\mathcal{A}}$ runs on input U, W, aux, r in time $t_{\mathcal{E}}$ and outputs $\hat{Z}_1, \hat{Z}_2 \in G$ or $\perp$ such that

- If $\mathcal{E}^{DH}_{\mathcal{A}}(U, W, \mathsf{aux}, r) \neq \perp$ then $\hat{Z}_1 = DH(U, Z)$, $\hat{Z}_2 = DH(W, Z)$
- For every algorithm C running in time $t_{\mathcal{D}}$, we have that $$\Pr[C(U, W, Z, r, \mathsf{aux}) \in \{DH(U, Z), DH(W, Z)\} \mid \mathcal{E}^{DH}_{\mathcal{A}}(U, W, \mathsf{aux}, r) = \perp] \leq \epsilon_{\mathcal{D}},$$ where $Z = \mathcal{A}(U, W, \mathsf{aux}, r)$ and the probability is taken over the coins of C and uniform distribution on (U, W, r).

B.2 Key Encapsulation Mechanisms (KEMs)

We briefly review the syntax for KEMs.

Definition B.5 (Key Encapsulation Mechanisms). A *key encapsulation mechanism (KEM)* is a triple of algorithms KEM = (KGen, Enc, Dec) with associated ciphertext space C and key space $\mathcal{K}$. In more detail:

- $\mathsf{KGen}() \xrightarrow{\$} (\mathsf{pk}, \mathsf{sk})$: A probabilistic algorithm that outputs a key pair (pk, sk)
- $\mathsf{Enc}(\mathsf{pk}) \xrightarrow{\$} (ct, ss)$: A probabilistic algorithm taking as input a public key pk and outputs a ciphertext $ct \in C$ and the therein encapsulated key $ss \in \mathcal{K}$.
- $\mathsf{Dec}(\mathsf{sk}, ct) \to ss'$: A deterministic algorithm taking as input a ciphertext $ct \in C$ and secret key sk and outputs $ss \in \mathcal{K} \cup \{\perp\}$, where $\perp$ indicates an error.

We define the key collision probability γ_{coll} for a KEM KEM as the probability of the public keys of two independently sampled key pairs to coincide.

We say that KEM = (KGen, Enc, Dec) is δ-*correct* if, for every key pair $(\mathsf{pk}, \mathsf{sk}) \xleftarrow{\$} \mathsf{KGen}()$ and every encapsulation $(ct, ss) \xleftarrow{\$} \mathsf{Enc}(\mathsf{pk})$, it holds that $\Pr[ss' \neq ss \mid ss' \leftarrow \mathsf{Dec}(\mathsf{sk}, ct)] \leq \delta$.

B.3 Authenticated Encryption with Associated Data (AEAD)

We follow the exposition of [50].

Definition B.6 (AEAD). An Authenticated Encryption scheme with Associated Data is a pair of algorithms AEAD = (Enc, Dec) with associated key space $\mathcal{K}$, nonce space $\mathcal{N}$, associated data (header) space $\mathcal{H}$, message space $\mathcal{M}_{\mathsf{AE}}$, and ciphertext space C, defined as follows:

- $\mathsf{Enc}(k, n, AD, \mu) \to ct$: On input a key k, a nonce n, associated data AD, and a message μ, this deterministic algorithm returns a ciphertext ct.
- $\mathsf{Dec}(k, n, AD, ct) \to \mu$: On input a key k, a nonce n, associated data AD, and a ciphertext ct, this deterministic algorithm returns a message μ or a distinguished error symbol $\perp \notin \mathcal{M}_{\mathsf{AE}}$.

```
ChallPassive($U, V, \text{info}_{\text{create}}, \vec{\mu}$):
$\pi_U$.oid $\leftarrow U$;    $\pi_U$.pid $\leftarrow V$
$\pi_V$.oid $\leftarrow V$;    $\pi_V$.pid $\leftarrow U$
$\pi_U$.role $\leftarrow$ initiator;    $\pi_V$.role $\leftarrow$ responder
$m_1 \leftarrow (\text{create}, \text{info}_{\text{create}})$
$(\mu_1, \ldots, \mu_{n_m}) \leftarrow \vec{\mu}$
for $i \in [1, 3, \ldots, n_m - 1]$ //until $n_m$ if $n_m$ is odd
  if $b = 0$
    $(\pi_U, m_{i+1}) \leftarrow\$ \text{Run}(\text{sk}_U, \vec{\text{pk}}, \pi_U, m_i, \mu_i)$
    $(\pi_V, m_{i+2}) \leftarrow\$ \text{Run}(\text{sk}_V, \vec{\text{pk}}, \pi_V, m_{i+1}, \mu_{i+1})$
  else
    $(\pi_U, m_{i+1}) \leftarrow\$ \text{Fake}^{\text{O}_F}(\vec{\text{pk}}, \pi_U, m_i, \mu_i, \text{aux})$
    $(\pi_U, m_{i+2}) \leftarrow\$ \text{Fake}^{\text{O}_F}(\vec{\text{pk}}, \pi_U, m_{i+1}, \mu_{i+1}, \text{aux})$ //all Fake invocations share a state
$Q[\pi_U] \leftarrow (m_i)_{i=1}^{nm}$;    $K[\pi_U] \leftarrow \pi_U.K$
return $(Q[\pi_U], K[\pi_U])$
```

Figure 6: The challenge oracle for passive adversaries.

We say that an AEAD scheme is *correct* if, for every $k \in \mathcal{K}$, $n \in \mathcal{N}, AD \in \mathcal{H}, \mu \in \mathcal{M}_{\mathsf{AE}}$ it holds that $\Pr[\mathsf{Dec}(k, n, AD, \mathsf{Enc}(k, n, AD, \mu)) = \mu] = 1$.

Note that throughout the paper we omit the nonce since Signal's implementations of X3DH and PQXDH derive the nonce deterministically from the key.

C EXTENDING THE MODEL TO PASSIVE ADVERSARIES

For the sake of completeness, we extend our model to passive adversaries by adding a ChallPassive oracle, which the adversary may have access to via O_A, given in Figure 6.

In consequence, we also need to adapt Definition 3.1 to allow ChallPassive $\in \text{O}_A$. We obtain *deniability against passive adversaries* by setting $\{\text{ChallPassive}\} = \text{O}_A$, allowing at most q_C queries to the ChallPassive oracle. The ChallPassive limits the adversary to passively observing transcripts, without the opportunity to actively participate in the protocol execution or learning a party's secret keys.

It is easy to see that deniability against semi-honest adversaries implies deniability against passive adversaries.

D DEFERRED PROOFS FOR X3DH

Here we give the deferred proofs for initiator deniability of X3DH. We show responder deniability for both protocols together in Appendices E.3 and E.4. Table 2 gives an overview over all theorems.

D.1 Proof of Theorem 4.1

Proof. We give the Fake algorithm in Figure 4 on the left-hand side and Figure 7 on the left-hand side.

The Fake algorithm simulates Bob's pre-key bundle by sampling a fresh ephemeral key (if needed) and learns a signature on the semi-static key from the auxiliary info aux. Furthermore, to process Alice's message, the Fake algorithm learns Alice's long-term secret key from the SK oracle and Alice's ephemeral secret key esk_A^{DH} from the randomness that was previously used to create Alice's message. Using these secret keys, it can compute DH_1 (with esk_A^{DH}), and DH_2, DH_3, DH_4 (with $ltsk_A^{\mathsf{DH}}$), and in consequence the session key. Using the session key, Fake decrypts the AEAD ciphertext. If decryption fails, it aborts.

Since Fake produces a transcript and session key in the same way as Run, the distinguisher cannot have an advantage in winning his game, even if the distinguisher has access to all long-term and semi-static secret keys. □

D.2 Proof of Theorem 4.2

This theorem adapts the proof of [56, Theorem 6] to our model. Since [56] considers only keys of two lifetimes (which they call long-term and ephemeral), we extrapolate to long-term, semi-static, and ephemeral keys. Their ephemeral key gets signed, just like the semi-static key in our description of X3DH. And if their ephemeral keys were used only once, then the simulator using a signature from the auxiliary info would raise the distinguisher's suspicion (since the distinguisher knows the auxiliary information as well). Hence, we consider their "ephemeral" keys as semi-static. Furthermore, they do not consider the AEAD ciphertext.

The proof of [56] relies on key registration: They argue that at the time of key registration a user needs to provide an extractable proof of knowledge of the secret key. Hence, they provide the long-term secret key of the adversary to the simulator. We model this by limiting the adversary to honestly generated keys with the RegHon oracle and by giving the Fake algorithm access to the SK oracle.[16]

Proof. We give the Fake algorithm in Figure 4 on the left-hand side and Figure 7.

The Fake algorithm uses Alice's secret key from the SK oracle to compute DH_1, Bob's ephemeral secret, which it previously sampled, to compute DH_4, and the extractor from the K2DH assumption to learn DH_2 and DH_3. If the extraction fails, Fake sets a random key as session key. Using the session key, Fake decrypts the AEAD ciphertext. If decryption fails, it aborts.

Formally, the adversary $\mathcal{A}$ never outputs any value. Here, we apply the extractor $\mathcal{E}_{\mathcal{A}}^{DH}$ in line 70 to $\mathcal{A}$ querying the Chall oracle with the message $(A, epk_A^{\mathsf{DH}}, ct_{\mathsf{AE}})$ after receiving a message from Bob with semi-static key ssid.

Since Fake produces the transcript in the same way as Run, it does not help the distinguisher to win. If the extractor succeeds, then Fake has computed the session key in the same way as Run, allowing no advantage for the distinguisher. If the extractor fails, then the distinguisher succeeds in extracting the DH_2 or DH_3 with a maximum probability of $\epsilon_{\mathcal{D}^{DH}}$. Hence, the distinguisher can distinguish the real session key from the simulated session key with a probability of at most $\epsilon_{\mathcal{D}^{DH}}$. □

D.3 Proof of Theorem 4.3

Proof. We give the Fake algorithm in Figure 4 on the left-hand side and Figure 5.

The Fake algorithm simulates Bob's pre-key bundle by sampling a fresh ephemeral key (if needed) and learns a signature on the semi-static key from the auxiliary info aux. Furthermore, to process Alice's message the Fake algorithm uses two techniques: It relies on the extractor from the EKDH assumption to learn DH_1 through DH_3 and uses Bob's ephemeral key (from its previous run) to compute DH_4. If any of the three extractions fail, Fake sets a random key

[16]These oracles also return semi-static secret keys, which we deem in spirit consistent with the idea of [56].

```
Fake($\vec{\mathsf{pk}}, \pi, m = (A, epk_A^{\mathsf{DH}}, ct_{\mathsf{AE}}, ct_{\mathsf{KEM}}), \mu, \mathsf{aux}, r_C$):
47 $\pi.\mathsf{pid} \leftarrow A$
48 $(ltsk_A^{\mathsf{DH}}, ltsk_A^{\mathsf{SIG}}) \leftarrow$ from SK($\pi$)
49 $esk_A^{\mathsf{DH}} \leftarrow$ from $r_C$
50 $(sspk_B^{\mathsf{DH}}, sspk_B^{\mathsf{KEM}}) \leftarrow sspk_B^{\mathsf{ssid}}$
51 $DH_1 \leftarrow \mathsf{DH}(ltsk_A^{\mathsf{DH}}, sspk_B^{\mathsf{ssid}})$
52 $DH_2 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, ltpk_B^{\mathsf{DH}})$
53 $DH_3 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, sspk_B^{\mathsf{ssid}})$
54 if $epk_B \neq \bot$ //full handshake
55     $DH_4 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, epk_B^{\mathsf{DH}})$
56 else //reduced handshake
57     $DH_4 \leftarrow \epsilon$
58 $ss \leftarrow$ from $r_C$ //encapsulation against $epk_B^{\mathsf{KEM}}$ or $sspk_B^{\mathsf{KEM}}$
59 $\mathsf{ms} \leftarrow DH_1 \| DH_2 \| DH_3 \| DH_4 \| ss$
60 $\pi.\mathsf{K} \leftarrow \mathsf{KDF}(\mathsf{ms})$
61 $AD \leftarrow ltpk_A \| ltpk_B$
62 $\mu' \leftarrow \mathsf{Dec}(\pi.\mathsf{K}, AD, ct_{\mathsf{AE}})$
63 if $\mu' = \bot$ then $\pi.\mathsf{K} \leftarrow \bot$
64 return $(\pi, \epsilon, \mu')$
```

```
Fake($\vec{\mathsf{pk}}, \pi, m = (A, epk_A^{\mathsf{DH}}, ct_{\mathsf{AE}}, ct_{\mathsf{KEM}}), \mu, \mathsf{aux}, r$):
65 $\pi.\mathsf{pid} \leftarrow A$
66 $(ltsk_A^{\mathsf{DH}}, ltsk_A^{\mathsf{SIG}}) \leftarrow$ extract from SK($\pi$)
67 $esk_B^{\mathsf{DH}} \leftarrow$ from previous run
68 $(sspk_B^{\mathsf{DH}}, sspk_B^{\mathsf{KEM}}) \leftarrow sspk_B^{\mathsf{ssid}}$
69 $DH_1 \leftarrow \mathsf{DH}(ltsk_A^{\mathsf{DH}}, sspk_B^{\mathsf{ssid}})$
70 $(DH_2, DH_3) \leftarrow \mathcal{E}_{\mathcal{A}}^{DH}(ltpk_B^{\mathsf{DH}}, sspk_B^{\mathsf{ssid}}, \mathsf{aux}, r)$
71 if $epk_B \neq \bot$ //ephemeral DH key present
72     $DH_4 \leftarrow \mathsf{DH}(epk_A^{\mathsf{DH}}, esk_B^{\mathsf{DH}})$
73 else
74     $DH_4 \leftarrow \epsilon$
75 if $epk_B^{\mathsf{KEM}} \neq \bot$ //ephemeral KEM key present
76     $ss \leftarrow\$ \mathcal{E}_{\mathcal{A}^{\mathsf{O}_A}}^{PA}(epk_B^{\mathsf{KEM}}, ct_{\mathsf{KEM}}, r)$
77 else //no ephemeral KEM key present
78     $ss \leftarrow\$ \mathcal{E}_{\mathcal{A}^{\mathsf{O}_A}}^{PA}(sspk_B^{\mathsf{KEM}}, ct_{\mathsf{KEM}}, r)$
79 if $DH_2 = \bot \vee DH_3 = \bot$
80     $\pi.\mathsf{K} \leftarrow\$ \{0,1\}^{256}$
81 else
82     $\mathsf{ms} \leftarrow DH_1 \| DH_2 \| DH_3 \| DH_4 \| ss$
83     $\pi.\mathsf{K} \leftarrow \mathsf{KDF}(\mathsf{ms})$
84 $AD \leftarrow ltpk_A \| ltpk_B$
85 $\mu' \leftarrow \mathsf{Dec}(\pi.\mathsf{K}, AD, ct_{\mathsf{AE}})$
86 if $\mu' = \bot$ then $\pi.\mathsf{K} \leftarrow \bot$
87 return $(\pi, \epsilon, \mu')$
```

Figure 7: The Fake **algorithm processing Alice's message (initiator deniability) in Theorems 4.1 and 4.5 (on the left-hand side, Appendices D.1 and E.1) and Theorems 4.2 and 4.6 (on the right-hand side, Appendices D.2 and E.2).**

as session key; otherwise it computes the session key normally. It decrypts the AEAD ciphertext using the session key and erases the session key in case decryption fails.

Formally, the adversary $\mathcal{A}$ never outputs any value. Here, we apply the extractor $\mathcal{E}_{\mathcal{A}}^{DH}$ to those cases were $\mathcal{A}$ queries some oracle with a message: In line 26, the adversary's output refers to $\mathcal{A}$ querying the Reg oracle (with a long-term and a semi-static public key, but we only care about the long-term key); in lines 27 and 28, the adversary's output refers to $\mathcal{A}$ querying the ChallInit oracle on message $(A, epk_A^{\mathsf{DH}}, ct_{\mathsf{AE}})$.

The Fake algorithm produces the transcript in the same way as Run and, hence, the transcript is indistinguishable. If the extractor succeeds in all three cases, then the Fake algorithm has computed the session key in the same way as Run, i.e. indistinguishably. If the extractor fails in any of the three cases, then the distinguisher succeeds in extracting one of the shared DH secrets with a maximum probability of $\epsilon_{\mathcal{D}^{DH}}$ as per the EKDH assumption. Hence, the distinguisher can distinguish the real session key from the simulated session key with a probability of at most $\epsilon_{\mathcal{D}^{DH}}$. Note that the second and third extractor call refer to the same output of the adversary, i.e. they are not independent. Conservatively, we bound with $\epsilon_{\mathcal{D}^{DH}}$ and not $(\epsilon_{\mathcal{D}^{DH}})^3$. □

E DEFERRED PROOFS FOR PQXDH

Here we give the deferred proofs for initiator deniability of PQXDH, and then for responder deniability of both protocols. Table 2 gives an overview over all theorems.

E.1 Proof of Theorem 4.5

Proof. We give the Fake algorithm in Figure 4 on the left-hand side and Figure 7.

The Fake algorithm simulates Bob's pre-key bundle by sampling a fresh ephemeral DH key (if needed) and learns a signature on the semi-static DH key and on the KEM key from the auxiliary info aux. Note that the Fake algorithm may need a signature for any semi-static KEM key or a signed ephemeral key for each of the q_{CI} queries. To process Alice's message, the Fake algorithm computes the DH shared secrets as for Theorem 4.1 (using Alice's freshly sampled ephemeral secret and the long-term secret obtained from the SK oracle). Furthermore, Fake extracts the KEM shared secret from the randomness that was previously used for the encapsulation. It computes the session key honestly as well as decrypts the AEAD ciphertext and aborts if decryption fails.

Since Fake produces a transcript and session key in the same way as Run and the distinguisher cannot detect reuse of the signature on the ephemeral KEM key, the distinguisher cannot have an advantage in winning the game, even if the distinguisher has access to all long-term and semi-static secret keys. □

E.2 Proof of Theorem 4.6

Proof. We give the Fake algorithm in Figure 4 on the left-hand side and Figure 7.

The Fake algorithm simulates Bob's pre-key bundle by sampling a fresh ephemeral DH key (if needed) and learns a signature on the semi-static DH key and on the KEM key from the auxiliary info aux. Note that the Fake algorithm may need a signature for any semi-static KEM key or a signed ephemeral key for each of the q_{CI} queries. When processing Alice's message, Fake uses Alice's long-term secret key from the SK oracle and the previously sampled ephemeral key to compute DH_1 and DH_4. It extracts DH_2, DH_3 from the adversary under the K2DH assumption. Furthermore, it uses the PA1 extractor in line 76 or 78 (depending on whether an ephemeral or semi-static KEM key is used) to learn the KEM shared secret. In contrast, Run computes DH_2, DH_3 with Bob's DH secret keys and the KEM shared secret ss with KEM.Dec() and Bob's corresponding KEM secret key. If the K2DH extraction fails, Fake sets a random key as session key. Else, it computes the session key honestly. It decrypts the AEAD ciphertext and aborts if decryption fails.

In a nutshell, the Fake algorithm creates Bob's pre-key bundle virtually identical to Run.

We argue over a series of game hops that the distinguisher cannot distinguish whether Alice's message was processed by Run or Fake.

Game 0 = Run. The initial game is the original deniability game $\mathcal{G}^{O_A,O_F,O_D\text{-}den}_{\mathsf{PQXDH},\mathsf{AuxPrep},n_p,n_{ss}}(\mathcal{A},\mathcal{D})$ with challenge $b = 0$, executing Run.

Game 1 (SK oracle). In this game Fake obtains Bob's long-term and semi-static secret key using the SK oracle, instead of receiving Bob's secret keys as input. This change is purely syntactical and not noticeable to an attacker or distinguisher. Hence, the winning probabilities for Games 0 and 1 are identical.

Game 2 (PA1 extractor). We replace the honest decapsulation of the KEM ciphertext with the PA extractor in lines 76 and 78. (In each invocation of Fake for processing Alice's message exactly one of these two lines is executed, depending on e_{DH}.)

We bound the advantage difference introduced by this step by the $(n_k, t_C, t_{\mathcal{E}^{PA}}, t_{\mathcal{D}^{PA}}, \epsilon_{\mathcal{D}^{PA}})$-PA1 assumption. We consider the reduction $\mathcal{B}$ – running the adversary with its oracles, excluding calls to the Decaps oracle – as ciphertext creator C of the PA1 game. The reduction $\mathcal{B}$ is started with a list of KEM keys $\vec{\mathsf{pk}}_{\mathsf{KEM}}$ and explicit randomness $r_C = (r_{\mathcal{A}}, r_{RH}, r_C)$.

The reduction samples $n_p + n_p \cdot n_{ss}$ DH keys and sorts them by user into $\vec{\mathsf{pk}}$. Next, it prepares the auxiliary info aux, which consists of one pre-key bundle per user and semi-static key and q_{CI} pre-key bundles per user with ephemeral KEM keys. Note that for both sampling the DH keys and preparing the auxiliary info the reduction uses hard-wired randomness. (Without hard-wiring the randomness the PA1 extractor would require this randomness as input since we view the reduction as ciphertext creator.) Then, it starts the adversary $\mathcal{A}$ on the auxiliary info aux and the list of public keys $\vec{\mathsf{pk}}$ with the randomness $r_{\mathcal{A}}$. The reduction provides the adversary with access to its oracles $O_A = \{\text{RegHon}, \text{Init}, \text{ChallInit}\}$. To answer queries to the RegHon oracle the reduction uses the randomness r_{RH}. To answer queries to the ChallInit oracle the reduction always uses Fake as described for the previous game with the following three changes: First, instead of sampling a new ephemeral KEM key for Bob's pre-key bundle, it uses the next KEM public key from its input list. Second, the reduction replaces lines 76 and 78 (the decapsulation or extraction of the KEM shared secrets) with a call to its own Decaps oracle of the PA1 game. Third, it uses the randomness r_C where needed. After the adversary terminates, the reduction outputs $(\mathsf{aux}, \vec{\mathsf{pk}}, r, Q, K)$. The PA1-distinguisher $\mathcal{D}'$ then runs the deniability-distinguisher $\mathcal{D}$ on $(\mathsf{aux}, \vec{\mathsf{pk}}, r, Q, K)$ and returns the same guess as the deniability-distinguisher $\mathcal{D}$.

Hence, depending on the secret bit of the PA1 game, the reduction simulates either the previous game or the current game. We include the running time of the $O_A = \{\text{RegHon}, \text{Init}, \text{Chall}\}$ oracles in $t_{\mathcal{A}}$, allowing for $t_{\mathcal{A}} \approx t_C$.

The PA1 distinguisher $\mathcal{D}'$ executes the deniability distinguisher $\mathcal{D}$, resulting in essentially the same run time $t_{\mathcal{D}} \approx t_{\mathcal{D}^{PA}}$, and directly returns the guess of the deniability distinguisher $\mathcal{D}$. If the deniability distinguisher $\mathcal{D}$ succeeds, then the PA1 distinguisher $\mathcal{D}'$ succeeds as well and can distinguish between the current and the previous game. By the PA1 assumption, the PA1 distinguisher succeeds with an advantage $\leq \epsilon_{\mathcal{D}^{PA}}$, which also limits the advantage of the deniability distinguisher.

Game 3 = Fake *(K2DH assumption).* We replace the honest computations of DH_2, DH_3 with the K2DH extractor in line 70, finally resulting in Fake as described in Figure 7 on the right-hand side. We bound the advantage difference introduced by this step by the $(t_{\mathcal{A}^{DH}}, t_{\mathcal{E}^{DH}}, t_{\mathcal{D}^{DH}}, \epsilon_{\mathcal{D}^{DH}})$-K2DH assumption.

Formally, the adversary $\mathcal{A}$ never outputs any value. Here, we apply the extractor $\mathcal{E}^{DH}_{\mathcal{A}}$ in line 70 to $\mathcal{A}$ querying the Chall oracle with the message $(A, epk^{\mathsf{DH}}_A, ct_{\mathsf{AE}}, ct_{\mathsf{KEM}})$ after receiving a message from Bob with semi-static key ssid.

If the extractor succeeds, then Fake has computed the session key in the same way as Run, allowing no advantage in distinguishing the current game from the previous game. If the extractor fails, then Fake sets a random session key. The runtime of the attacker in the K2DH assumption limits the runtime of the deniability attacker, i.e. $t_{\mathcal{A}} \approx t_{\mathcal{A}^{DH}}$. Furthermore, any K2DH distinguisher running in time $t_{\mathcal{D}^{DH}}$ succeeds in extracting DH_2 or DH_3 with a maximum probability of $\epsilon_{\mathcal{D}^{DH}}$. Hence, the deniability distinguisher with the same limit in runtime can distinguish the previous game from the current game with a probability of at most $\epsilon_{\mathcal{D}^{DH}}$. □

E.3 Proof of Theorem 4.8

Proof. We give the Fake algorithm in Figure 8 on the left-hand side. The Fake algorithm considers the KEM parts (obtaining the KEM public key, verifying the signature on the KEM public key, encapsulating against the KEM public key, and sending a KEM ciphertext) only for PQXDH.

The Fake algorithm uses the freshly sampled ephemeral key esk^{DH}_A to compute DH_1 and Bob's long-term secret keys from the SK oracle $ltsk^{\mathsf{DH}}_B$ to compute DH_2, DH_3, DH_4. It learns the KEM shared secret from encapsulating against Bob's public key. If either signature (on the semi-static DH key or on the KEM key) does not verify, it aborts. Using those secrets, Fake computes the session key and encrypts the user message with the AEAD scheme under the session key.

Since Fake produces a transcript and session key in the same way as Run, the distinguisher cannot have an advantage in winning his game, even if the distinguisher has access to all long-term and semi-static secret keys. □

E.4 Proof of Theorem 4.9

Proof. We give the Fake algorithm in Figure 8 on the right-hand side. The Fake algorithm considers the KEM parts (obtaining the KEM public key, verifying the signature on the KEM public key, encapsulating against the KEM public key, and sending a KEM ciphertext) only for PQXDH.

The Fake algorithm uses its freshly sampled ephemeral secret key esk^{DH}_A to compute DH_2 through DH_4. It extracts DH_1 from the

```
$\mathsf{Fake}(\vec{\mathsf{pk}}, \pi, m = (B, \mathsf{ssid}, \sigma_{\mathsf{DH}}^{\mathsf{ssid}}, epk_B, \sigma_{\mathsf{KEM}}), \mu, \mathsf{aux}, r)$:
$\pi.\mathsf{pid} \leftarrow B$
$sssk_B^{\mathsf{ssid}} \leftarrow$ extract from $\mathsf{SK}(\pi)$
$(sssk_B^{\mathsf{DH}}, sssk_B^{\mathsf{KEM}}) \leftarrow sssk_B^{\mathsf{ssid}}$
$(epk_A^{\mathsf{DH}}, esk_A^{\mathsf{DH}}) \leftarrow\$ \mathsf{DH.KGen}()$
$(sspk_B^{\mathsf{DH}}, sspk_B^{\mathsf{KEM}}) \leftarrow sspk_B^{\mathsf{ssid}}$
$(epk_B^{\mathsf{DH}}, epk_B^{\mathsf{KEM}}) \leftarrow epk_B$
if $\mathsf{SIG.Vf}(ltpk_B^{\mathsf{SIG}}, sspk_B^{\mathsf{DH}}, \sigma_{\mathsf{DH}}) = \mathsf{false}$
    return $(\pi, \epsilon, \epsilon)$
$DH_1 \leftarrow \mathsf{DH}(ltpk_A^{\mathsf{DH}}, sssk_B^{\mathsf{DH}})$
$DH_2 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, ltpk_B^{\mathsf{DH}})$
$DH_3 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, sspk_B^{\mathsf{DH}})$
if $epk_B^{\mathsf{DH}} \neq \perp$ //ephemeral DH key present
    $DH_4 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, epk_B^{\mathsf{DH}})$
else
    $DH_4 \leftarrow \epsilon$
if $epk_B^{\mathsf{KEM}} \neq \perp$ //ephemeral KEM key present
    if $\mathsf{SIG.Vf}(ltpk_B^{\mathsf{SIG}}, epk_B^{\mathsf{KEM}}, \sigma_{\mathsf{KEM}}) = \mathsf{false}$
        return $(\pi_A, \epsilon, \epsilon)$
    $(ct_{\mathsf{KEM}}, ss) \leftarrow\$ \mathsf{KEM.Enc}(epk_B^{\mathsf{KEM}})$
else //no ephemeral KEM key present
    if $\mathsf{SIG.Vf}(ltpk_B^{\mathsf{SIG}}, sspk_B^{\mathsf{KEM}}, \sigma_{\mathsf{KEM}}) = \mathsf{false}$
        return $(\pi_A, \epsilon, \epsilon)$
    $(ct_{\mathsf{KEM}}, ss) \leftarrow\$ \mathsf{KEM.Enc}(sspk_B^{\mathsf{KEM}})$
$\mathsf{ms} \leftarrow DH_1 \| DH_2 \| DH_3 \| DH_4 \| ss$
$\pi.\mathsf{K} \leftarrow \mathsf{KDF}(\mathsf{ms})$
$AD \leftarrow ltpk_A \| ltpk_B$
$ct_{\mathsf{AE}} \leftarrow \mathsf{Enc}(\pi_A.\mathsf{K}, AD, \mu)$
return $(\pi, (A, epk_A^{\mathsf{DH}}, ct_{\mathsf{AE}}, ct_{\mathsf{KEM}}), \epsilon)$
```

```
$\mathsf{Fake}(\vec{\mathsf{pk}}, \pi, m = (B, \mathsf{ssid}, \sigma_{\mathsf{DH}}^{\mathsf{ssid}}, epk_B, \sigma_{\mathsf{KEM}}), \mu, \mathsf{aux}, r)$:
$\pi.\mathsf{pid} \leftarrow B$
$(epk_A^{\mathsf{DH}}, esk_A^{\mathsf{DH}}) \leftarrow\$ \mathsf{DH.KGen}()$
$(sspk_B^{\mathsf{DH}}, sspk_B^{\mathsf{KEM}}) \leftarrow sspk_B^{\mathsf{ssid}}$
$(epk_B^{\mathsf{DH}}, epk_B^{\mathsf{KEM}}) \leftarrow epk_B$
if $\mathsf{SIG.Vf}(ltpk_B^{\mathsf{SIG}}, sspk_B^{\mathsf{DH}}, \sigma_{\mathsf{DH}}) = \mathsf{false}$
    return $(\pi, \epsilon, \epsilon)$
$DH_1 \leftarrow \mathcal{E}_{\mathcal{A}}^{DH}(ltpk_A^{\mathsf{DH}}, \mathsf{aux}, r)$
$DH_2 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, ltpk_B^{\mathsf{DH}})$
$DH_3 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, sspk_B^{\mathsf{DH}})$
if $epk_B^{\mathsf{DH}} \neq \perp$ //ephemeral DH key present
    $DH_4 \leftarrow \mathsf{DH}(esk_A^{\mathsf{DH}}, epk_B^{\mathsf{DH}})$
else
    $DH_4 \leftarrow \epsilon$
if $epk_B^{\mathsf{KEM}} \neq \perp$ //ephemeral KEM key present
    if $\mathsf{SIG.Vf}(ltpk_B^{\mathsf{SIG}}, epk_B^{\mathsf{KEM}}, \sigma_{\mathsf{KEM}}) = \mathsf{false}$
        return $(\pi_A, \epsilon, \epsilon)$
    $(ct_{\mathsf{KEM}}, ss) \leftarrow\$ \mathsf{KEM.Enc}(epk_B^{\mathsf{KEM}})$
else //no ephemeral KEM key present
    if $\mathsf{SIG.Vf}(ltpk_B^{\mathsf{SIG}}, sspk_B^{\mathsf{KEM}}, \sigma_{\mathsf{KEM}}) = \mathsf{false}$
        return $(\pi_A, \epsilon, \epsilon)$
    $(ct_{\mathsf{KEM}}, ss) \leftarrow\$ \mathsf{KEM.Enc}(sspk_B^{\mathsf{KEM}})$
if $DH_1 = \perp$
    $\pi.\mathsf{K} \leftarrow\$ \{0,1\}^{256}$
else
    $\mathsf{ms} \leftarrow DH_1 \| DH_2 \| DH_3 \| DH_4 \| ss$
    $\pi.\mathsf{K} \leftarrow \mathsf{KDF}(\mathsf{ms})$
$AD \leftarrow ltpk_A \| ltpk_B$
$ct_{\mathsf{AE}} \leftarrow \mathsf{Enc}(\pi_A.\mathsf{K}, AD, \mu)$
return $(\pi, (A, epk_A^{\mathsf{DH}}, ct_{\mathsf{AE}}, ct_{\mathsf{KEM}}), \epsilon)$
```

Figure 8: The Fake algorithms for producing Alice's message (responder deniability) in Theorem 4.8 (on the left-hand side, Appendix E.3) and Theorem 4.9 (on the right-hand side, Appendix E.4).

adversary under the EKDH assumption. As before, we interpret the adversary's query to the Reg oracle as its output and apply the extractor $\mathcal{E}_{\mathcal{A}}^{DH}$ accordingly. It learns the KEM shared secret from encapsulating against Bob's public key. If either signature (on the semi-static DH key or on the KEM key) does not verify, it aborts. If the extraction fails, Fake sets a random key as session key. Otherwise, it computes the session key honestly. It encrypts the user message with the AEAD scheme under the session key.

The Fake algorithm produces the transcript in the same way as Run and, hence, indistinguishably. If the extractor succeeds, then Fake has computed the session key in the same way as Run, and, hence, indistinguishably. If the extractor fails, then the distinguisher succeeds in extracting the shared DH secret with a maximum probability of $\epsilon_{\mathcal{D}^{DH}}$. Hence, the distinguisher can distinguish the real session key from the simulated session key with a probability of at most $\epsilon_{\mathcal{D}^{DH}}$. □