Quantum non-malleability and authentication

Gorjan Alagic and Christian Majenz

QMATH, Department of Mathematical Sciences
University of Copenhagen

galagic@gmail.com majenz@math.ku.dk

Abstract. In encryption, non-malleability is a highly desirable property: it ensures that adversaries cannot manipulate the plaintext by acting on the ciphertext. In [6], Ambainis et al. gave a definition of non-malleability for the encryption of quantum data. In this work, we show that this definition is too weak, as it allows adversaries to "inject" plaintexts of their choice into the ciphertext. We give a new definition of quantum non-malleability which resolves this problem. Our definition is expressed in terms of entropic quantities, considers stronger adversaries, and does not assume secrecy. Rather, we prove that *quantum non-malleability implies secrecy*; this is in stark contrast to the classical setting, where the two properties are completely independent. For unitary schemes, our notion of non-malleability is equivalent to encryption with a two-design (and hence also to the definition of [6]).

Our techniques also yield new results regarding the closely-related task of quantum authentication. We show that "total authentication" (a notion recently proposed by Garg et al. [18]) can be satisfied with two-designs, a significant improvement over the eight-design construction of [18]. We also show that, under a mild adaptation of the rejection procedure, both total authentication and our notion of non-malleability yield quantum authentication as defined by Dupuis et al. [16].

1 Introduction

Background. In its most basic form, encryption ensures secrecy in the presence of eavesdroppers. Besides secrecy, another desirable property is *non-malleability*, which guarantees that an active adversary cannot modify the plaintext by manipulating the ciphertext. In the classical setting, secrecy and non-malleability are independent: there are schemes which satisfy secrecy but are malleable, and schemes which are non-malleable but transmit the plaintext in the clear. If both secrecy and non-malleability is desired, then pairwise-independent permutations provide information-theoretically perfect (one-time) security [20]. In the computational security setting, non-malleability can be achieved by MACs, and ensures chosen-ciphertext security for authenticated encryption.

In the setting of quantum information, encryption is the task of transmitting quantum states over a completely insecure quantum channel. Information-theoretic secrecy for quantum encryption is well-understood. Non-malleability,

on the other hand, has only been studied in one previous work, by Ambainis, Bouda and Winter [6]. Their definition (which we will call ABW-non-malleability, or ABW-NM) requires that the scheme satisfies secrecy, and that the "effective channel" $\mathsf{Dec} \circ \Lambda \circ \mathsf{Enc}$ of any adversary Λ amounts to either the identity map or replacement by some fixed state. In the case of unitary schemes, ABW-NM is equivalent to encrypting with a unitary two-design. Unitary two-designs are a natural quantum analogue of pairwise-independent permutations, and can be efficiently constructed in a number of ways (see, e.g., [10, 14].)

While quantum non-malleability has only been considered by [6], the closely-related task of quantum authentication (where decryption is allowed to reject) has received significant attention (see, e.g., [2, 7, 11, 16, 18].) The widely-adopted definition of Dupuis, Nielsen and Salvail asks that the averaged effective channel of any adversary is close to a map which does not touch the plaintext [16]; we refer to this notion as DNS-authentication. Recent work by Garg, Yuen and Zhandry [18] established another notion of quantum authentication, which they call "total authentication." The notion of total authentication has two major differences from previous definitions: (i.) it asks for success with high probability over the choice of keys, rather than simply on average, and (ii.) it makes no demands whatsoever in the case that decryption rejects. We refer to this notion of quantum authentication as GYZ-authentication. In [18], it is shown that GYZ-authentication can be satisfied with unitary eight-designs.

This work. In this work, we devise a new definition of non-malleability (denoted NM) for quantum encryption, improving on ABW-NM in a number of ways. First, our definition is expressed in terms of entropic quantities, which allows us to bring several quantum-information-theoretic techniques to bear (such as decoupling.) Second, we consider more powerful adversaries, which can possess side information about the plaintext. Third, we remove the possibility of a "plaintext injection" attack, whereby an adversary against an ABW-NM scheme can send a plaintext of their choice to the receiver. Finally, our definition does not demand secrecy; instead, we show that *quantum secrecy is a consequence of quantum non-malleability.* This is a significant departure from the classical case, and is analogous to the fact that quantum authentication implies secrecy [7].

The primary consequence of our work is twofold: first, encryption with unitary two-designs satisfies all of the above notions of quantum non-malleability; second, when equipped with blank "tag" qubits, the same scheme also satisfies all of the above notions of quantum authentication. A more detailed summary of the results is as follows. For schemes which have unitary encryption maps, we prove that NM is equivalent to encryption with unitary two-designs, and hence also to ABW-NM. For non-unitary schemes, we prove a characterization theorem for NM schemes that shows that NM implies ABW-NM, and provide a strong separation example between NM and ABW-NM (the aforementioned plaintext injection attack). In the case of GYZ authentication, we prove that two-designs (with tags) are sufficient, a significant improvement over the state-of-the-art, which requires eight-designs [18]. Moreover, the simulation of adversaries in this

proof is efficient, in the sense of Broadbent and Wainewright [11]. Finally, we show that GYZauthentication implies DNS-authentication, and that equipping an arbitrary NM scheme with tags yields DNS-authentication.

We remark that, after the initial version of our results was submitted, an independent work of C. Portmann gave an alternative proof that GYZ-authentication can be satisfied by the 2-design scheme [26].

1.1 Summary of contributions

In the following, all schemes are symmetric-key encryption schemes for quantum data, in the information-theoretic security setting.

Quantum non-malleability. We begin with non-malleability, in both the perfect setting (Section 3) and the approximate setting (Section 4).

1. **New definition of non-malleability.** We give a new definition of quantum non-malleability (NM), in terms of the information gain of an adversary's *effective attack* on the plaintext. The quantum registers are: plaintext A, ciphertext C, user's reference R, and adversary's side information B.

   Definition 1.1 (NM, informal) *A scheme is non-malleable (NM) if for any ϱ_{ABR} and any attack $\Lambda_{CB\to C\tilde{B}}$, the effective attack $\tilde{\Lambda}_{AB\to A\tilde{B}}$ satisfies*

   $$I(AR:\tilde{B})_{\tilde{\Lambda}(\varrho)} \leq I(AR:B)_{\varrho} + h(p_{=}(\Lambda,\varrho)).$$

   The binary entropy term is necessary because adversaries can always simply record whether they disturbed the ciphertext (see Definition 3.4).
2. **Results on non-malleability.** Our first result is an alternative characterization of NM, in terms of the form of the effective map $\tilde{\Lambda}$.

   Theorem 1.2 (informal) *A scheme is NM if and only if, for any attack $\Lambda_{CB\to C\tilde{B}}$, there exist maps $\Lambda'_{B\to\tilde{B}}$, $\Lambda''_{B\to\tilde{B}}$ such that the effective attack satisfies*

   $$\tilde{\Lambda} = \mathrm{id}_A \otimes \Lambda' + \frac{1}{|C|^2-1}\left(|C|\,\langle D_K(\mathbb{1}_C)\rangle - \mathrm{id}\right)_A \otimes \Lambda''.$$

   The fact that NM implies ABW-NM is an immediate corollary. The new definition is strictly stronger than ABW-NM: we give a scheme which is secure under ABW-NM but insecure under NM. This scheme is in fact susceptible to a powerful attack, whereby a simple adversary can replace the output of decryption with a plaintext of the adversary's choice. On the other hand, if we restrict our attention to schemes where the encryption maps are unitary, then we are able to show the following.

   Theorem 1.3 (informal) *Let Π be a scheme such that encryption E_k is unitary for all keys k. Then Π is NM if and only if $\{E_k\}_k$ is a two-design.*

   By the results of [6], we conclude that NM and ABW-NM are in fact equivalent for unitary schemes. Finally, we show that NM implies secrecy.

Theorem 1.4 (informal) *Quantum non-malleability implies secrecy.*

3. **Authentication from non-malleability.** Our final result in the setting of non-malleability shows that, by adding a "tag" space to the plaintext (as in the Clifford scheme [2]), we can turn an NM scheme into an authentication scheme as defined in [16]. More precisely, given an encryption scheme $\Pi = \{E_k\}$, we define Π_t^{tag} to be a new scheme whose encryption is $\varrho \mapsto E_k\big(\varrho_A \otimes |0\rangle\langle 0|_B^{\otimes t}\big)E_k^\dagger$, and whose decryption rejects unless B measures to $|0^t\rangle$.

   Theorem 1.5 (informal) *Let $\Pi = \{E_k\}$ be an encryption scheme. If Π is NM, then Π_t^{tag} is 2^{2-t}-DNS-authenticating.*

Quantum authentication. Our results on quantum authentication are summarized as follows. We note that, strictly speaking, our definitions of authentication deviate slightly from the original versions [16, 18], in that decryption outputs a reject symbol in place of the plaintext (rather than setting an auxiliary bit to "reject.") This adaptation is convenient for reasons we will return to later.

1. **GYZ implies DNS.** First, we show that GYZ-authentication implies DNS-authentication. We remark that this is not trivial: on one hand, GYZ strengthens DNS by requiring high probability of success (rather than success on-average); on the other hand, in the reject case GYZ requires nothing while DNS makes rather stringent demands. Nonetheless, we show the following.

   Theorem 1.6 (informal) *Let Π be an encryption scheme. If Π is ε-GYZ-authenticating, then it is also $O(\sqrt{\varepsilon})$-DNS-authenticating.*

2. **GYZ is achievable with 2-designs.** Next, we show that GYZ-authenticationis achieved with a "tagged" two-design scheme. The analysis of [18] required eight-designs for the same construction.

   Theorem 1.7 (informal) *Let $\Pi = \{E_k\}_k$ be a 2^{-t}-approximate 2-design scheme. Then Π_t^{tag} is $2^{-\Omega(t)}$-GYZ-authenticating.*

3. **GYZ authentication from non-malleability.** As a straightforward consequence of Theorem 1.3 and Theorem 1.7, we finally record that tagging a unitary non-malleable scheme results in a GYZ-authenticating scheme.

   Corollary 1.8 (informal) *There exists a constant $r > 0$ such that the following holds. If Π is a unitary $\Omega(2^{-rn})$-NM scheme for n-qubit messages, and $t = \text{poly}(n)$, then Π_t^{tag} is $2^{-\Omega(\text{poly}(n))}$-GYZ-authenticating.*

   A sufficiently strong NM scheme can be constructed via the ϵ-approximate version of Theorem 1.3 (see Theorem 4.5 and Remark 2.3 below.)

The remainder of the paper is structured as follows. In Section 2, we review some basic facts regarding quantum states, registers, and channels, and recall several useful facts about unitary designs. In Section 3, we consider the

exact setting, beginning with perfect secrecy and then continuing to perfect non-malleability (NM) and the relevant new results; we also discuss the relationship to ABW-NM in detail. We continue in Section 4 with the approximate setting, again beginning with secrecy and then continuing to approximate non-malleability. We end with the new results on quantum authentication, in Section 4.2.

2 Preliminaries

2.1 Quantum states, registers, and channels.

We assume basic familiarity with the formalism of quantum states, operators, and channels. We denote quantum registers (i.e., systems and their subsystems) with capital Latin letters, e.g., A, B, C. The Hilbert space corresponding to system A is denoted by $\mathcal{H}_A$. For a register A, we denote the dimension of $\mathcal{H}_A$ by $|A|$. We emphasize that, in this work, all Hilbert spaces will be finite-dimensional.

The space operators on $\mathcal{H}_A$ is denoted $\mathcal{B}(\mathcal{H}_A)$. We say that a quantum state is classical if it is diagonal in the standard (i.e., computational) basis. We denote the adjoint of an operator $X \in \mathcal{B}(\mathcal{H})$ by $X^\dagger$ and its transpose with respect to the computational basis by X^T. Where necessary, we will write a quantum state $\varrho \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C)$ as ϱ_{ABC} to emphasize that the state is a multipartite state over registers A, B, and C. When such a state has already been defined, we will write reduced states by omitting the traced-out registers, e.g., $\varrho_A := \mathrm{Tr}_{BC}[\varrho_{ABC}]$. We single out some special states which will appear frequently. Fix two systems S, S' with $|S| = |S'|$. We let

$$|\phi^+\rangle_{SS'} = |S|^{-1/2} \sum_i |ii\rangle_{SS'} \qquad \text{and} \qquad \phi^+_{SS'} = |\phi^+\rangle\langle\phi^+|_{SS'}$$

denote the maximally entangled state on the bipartite system SS' (expressed as a pure state on the left, and as a density operator on the right.) Furthermore, we let $\Pi^-_{SS'} = \mathbb{1}_{SS'} - \phi^+_{SS'}$ and $\tau^-_{SS'} = \Pi^-_{SS'}/(|S|^2 - 1)$. We also set $\tau_S = \mathbb{1}_S/|S|$ to be the maximally mixed state on S.

We denote the von Neumann entropy of a state ϱ_A by $H(A)_\varrho$, and the joint entropy of ϱ_{AB} by $H(AB)_\varrho$. We recall that the quantum mutual information of ϱ_{AB} is defined by

$$I(A:B)_\varrho := H(A)_\varrho + H(B)_\varrho - H(AB)_\varrho \,.$$

The quantum conditional mutual information of ϱ_{ABC} is defined by

$$I(A:B|C)_\varrho := H(AC)_\varrho + H(BC)_\varrho - H(ABC)_\varrho - H(C)_\varrho \,.$$

These quantities are nonnegative [21] and satisfy a chain rule:

$$I(A:BC|D)_\varrho = I(A:B|D)_\varrho + I(A:C|BD)_\varrho \,.$$

We remark that the above also holds for trivial D. Together with the Stinespring dilation theorem [27], non-negativity [22] and the chain rule imply the data processing inequality

$$I(A:\tilde{B}|C)_{\Lambda(\varrho)} \leq I(A:B|C)_{\varrho}\,,$$

when Λ is a CPTP (completely-positive, trace-preserving) map from $\mathcal{B}(\mathcal{H}_B)$ to $\mathcal{B}(\mathcal{H}_{\tilde{B}})$. An important special case is where $B = B_1B_2$ and $\Lambda = \mathrm{Tr}_{B_2}$ discards the contents of B_2.

We will refer to valid transformations between quantum states as channels, or CPTP maps. We will sometimes also consider trace-non-increasing completely-positive (CP) maps. When necessary, we will emphasize the input and output spaces of a map $\Lambda : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_C)$ by writing $\Lambda_{AB\to C}$. We denote the identity channel on, e.g., register A by $\mathrm{id}_{A\to A}$ (or simply id_A) and the channel from register A to A' with constant output $\sigma_{A'}$ by $\langle\sigma\rangle_{A\to A'}$. When composing operators on many registers, and if the context allows, we will elide tensor products with the identity operator. So, for example, with Λ as above we may write $\tau_{CD} = \Lambda\,\varrho_{ABD}$ in place of $\tau_{CD} = (\Lambda \otimes \mathrm{id}_D)\varrho_{ABD}$.

A standard tool in this setting is the Choi-Jamiołkowski (CJ) isomorphism [12, 19]. Let $\Xi_{A\to B} : \mathcal{B}(\mathcal{H}_A) \to \mathcal{B}(\mathcal{H}_B)$ be a linear operator. Then its CJ matrix is defined as

$$(\eta_\Xi)_{BA'} = \Lambda_{A\to B}(\phi^+_{AA'}). \tag{2.1}$$

The linear operator mapping Ξ to η_Ξ is an isomorphism of vector spaces and η_Ξ is positive semidefinite iff Ξ is CP. Moreover $\Xi_{A\to B}$ is TP iff $(\eta_\Xi)_{A'} = \tau_A$. The inverse of the CJ isomorphism is given by the equation

$$\Xi_{A\to B}(X_A) = |A|\mathrm{Tr}_{A'}\left[X^T_{A'}\,(\eta_\Xi)_{BA'}\right]. \tag{2.2}$$

We denote the swap operator by $F : |i\rangle \otimes |j\rangle \mapsto |j\rangle \otimes |i\rangle$.

Lemma 2.1 (Swap trick [17]) *For matrices A and B, $\mathrm{Tr}[AB] = \mathrm{Tr}[FA\otimes B]$.*

We will make frequent use of the trace norm $\|\cdot\|_1$, the operator norm $\|\cdot\|_\infty$, and the diamond norm $\|\Lambda_{A\to B}\|_\diamond := \max_{\varrho_{AA'}} \|\Lambda_{A\to B} \otimes \mathrm{id}_{A'}(\varrho_{AA'})\|_1$; here the max is taken over all pure quantum states $\varrho_{AA'}$ and $\mathcal{H}_A \cong \mathcal{H}_{A'}$. Recall that the Hölder inequality for operators states that, for any two operators X and Y,

$$\mathrm{Tr}[XY] \leq \|XY\|_1 \leq \|X\|_1\|Y\|_\infty\,. \tag{2.3}$$

2.2 Unitary designs.

We now recall the definition of unitary t-design, and some relevant variants. We begin by considering three different types of "twirls."

1. For a finite subset $\mathrm{D} \subset \mathrm{U}(\mathcal{H})$ of the unitary group on some finite dimensional Hilbert space $\mathcal{H}$, let

$$\mathcal{T}^{(t)}_{\mathrm{D}}(X) = \frac{1}{|\mathrm{D}|}\sum_{U\in\mathrm{D}} U^{\otimes t} X \left(U^\dagger\right)^{\otimes t} \tag{2.4}$$

be the associated t-twirling channel. If we take the entire unitary group (rather than just a finite subset), then we get the Haar t-twirling channel

$$\mathcal{T}^{(t)}_{\mathsf{Haar}}(X) = \int U^{\otimes t} X \left(U^\dagger\right)^{\otimes t} \mathrm{d}U. \tag{2.5}$$

2. We define the U-$\overline{U}$ twirl with respect to finite $\mathrm{D} \subset \mathrm{U}(\mathcal{H})$ by

$$\overline{\mathcal{T}}_{\mathrm{D}}(X) = \frac{1}{|\mathrm{D}|} \sum_{U \in \mathrm{D}} \left(U \otimes \overline{U}\right) X \left(U \otimes \overline{U}\right)^\dagger. \tag{2.6}$$

   The analogous U-$\overline{U}$ Haar twirling channel is denoted by $\overline{\mathcal{T}}_{\mathsf{Haar}}$.
3. The third notion is called a channel twirl, and is defined in terms of U-$\overline{U}$-twirling. Given a channel Λ, let η_Λ be the CJ state of Λ. The channel twirl $\mathcal{T}^{ch}_{\mathrm{D}}(\Lambda)$ of Λ is defined to be the channel whose CJ state is $\overline{\mathcal{T}}_{\mathrm{D}}(\eta_\Lambda)$.

Next, we define the three corresponding notions of designs.

Definition 2.2 *Let $\mathrm{D} \subset \mathrm{U}(\mathcal{H})$ be a finite set. We define the following.*

- *If $\left\|\mathcal{T}^{(t)}_{\mathrm{D}} - \mathcal{T}^{(t)}_{\mathsf{Haar}}\right\|_\diamond \leq \delta$ holds, then D is a δ-approximate t-design.*
- *If $\left\|\overline{\mathcal{T}}_{\mathrm{D}} - \overline{\mathcal{T}}_{\mathsf{Haar}}\right\|_\diamond \leq \delta$ holds, then D is a δ-approximate U-$\overline{U}$-twirl design.*
- *If $\left\|\mathcal{T}^{ch}_{\mathrm{D}}(\Lambda) - T^{ch}_{\mathsf{Haar}}(\Lambda)\right\|_\diamond \leq \delta$ holds for all CPTP maps Λ, then D is a δ-approximate channel-twirl design.*

For all three of the above, the case $\delta = 0$ is called an "exact design" (or simply "design".) All three notions of design are equivalent in the exact case. In the approximate case they are still connected, but there are some nontrivial costs in the approximation quality (See [23], Lemma 2.2.14, and an additional easy lemma proven in the full version [3]).

It is well-known that ε-approximate t-designs on n qubits can be generated by random quantum circuits of size polynomial in n, t and $\log(1/\varepsilon)$ [10]. In particular, the size of these circuits is polynomial even for exponentially-small choices of ε. We emphasize this observation as follows.

Remark 2.3 *Fix a polynomial t in n. Then, for any $\varepsilon > 0$, a random n-qubit quantum circuit consisting of $\mathrm{poly}(n, \log(1/\varepsilon))$ gates (from a universal set) satisfies every notion of ϵ-approximate t-design in Definition 2.2.*

For exact designs, we point out two important constructions. First, the prototypical example of a unitary one-design on n qubits is the n-qubit Pauli group. For exact unitary two-designs, the standard example is the Clifford group, which is the normalizer of the n-qubit Pauli group. Alternatively, the Clifford group is generated by circuits from the gate set $\{H, P, \mathrm{CNOT}\}$. It is well-known that one can efficiently generate exact unitary two-designs on n-qubits by building appropriate circuits from this gate set, using $O(n^2)$ random bits [1, 14].

3 The zero-error setting

We begin with the zero-error. In the case of secrecy, zero-error means that schemes cannot leak any information whatsoever. In the case of non-malleability, zero-error means that the adversary cannot increase their correlations with the secret by even an infinitesimal amount (except by trivial means; see below.)

3.1 Perfect secrecy

We begin with a definition of symmetric-key quantum encryption. Our formulation treats rejection during decryption in a slightly different manner from previous literature.

Definition 3.1 (Encryption scheme) *A symmetric-key quantum encryption scheme (QES) is a triple (τ_K, E, D) consisting of a classical state $\tau_K \in \mathcal{B}(\mathcal{H}_K)$ and a pair of channels*

$$E : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_K) \longrightarrow \mathcal{B}(\mathcal{H}_C \otimes \mathcal{H}_K)$$
$$D : \mathcal{B}(\mathcal{H}_C \otimes \mathcal{H}_K) \longrightarrow \mathcal{B}((\mathcal{H}_A \oplus \mathbb{C}|\perp\rangle) \otimes \mathcal{H}_K)$$

satisfying $[D \circ E](\cdot \otimes |k\rangle\langle k|) = (\mathrm{id}_A \oplus 0_\perp) \otimes |k\rangle\langle k|$ for all k.

The Hilbert spaces $\mathcal{H}_A$, $\mathcal{H}_C$ and $\mathcal{H}_K$ are implicitly given by the triple (τ_K, E, D). The state $|\perp\rangle$ is an error flag that allows the decryption map to report an error. For notational convenience when dealing with these schemes, we set

$$E_k = E(\cdot \otimes |k\rangle\langle k|) \qquad E_K = \mathrm{Tr}_K E(\cdot \otimes \tau_K)$$
$$D_k = D(\cdot \otimes |k\rangle\langle k|) \qquad D_K = \mathrm{Tr}_K D(\cdot \otimes \tau_K)\,.$$

We will often slightly abuse notation by referring to decryption maps D_k as maps from C to A; in fact, the output space of D_k is really the slightly larger space $\bar{A} := A \oplus \mathbb{C}|\perp\rangle$.

It is natural to define secrecy in the quantum world in terms of quantum mutual information. However, instead of asking for the ciphertext to be uncorrelated with the plaintext as in the classical case, we ask for the ciphertext to be uncorrelated from any reference system.

Definition 3.2 (Perfect secrecy) *A QES (τ_K, E, D) satisfies information-theoretic secrecy (ITS) if, for any Hilbert space $\mathcal{H}_B$ and any $\varrho_{AB} \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$, setting $\sigma_{CBK} = E(\varrho_{AB} \otimes \tau_K)$ implies $I(C : B)_\sigma = 0$.*

We note that, for perfect ITS, adding side information is unnecessary: the definition already implies that the ciphertext is in product with *any* other system. In particular, if the adversary has some auxiliary system E in their possession, then $I(B : CE)_\sigma = I(B : E)_\sigma$. Several definitions of secrecy for symmetric-key quantum encryption have appeared in the literature, but the above formulation

appears to be new. It can be shown that ITS is equivalent to perfect indistinguishability of ciphertexts (IND). The latter notion is a special case of an early indistinguishability-based definition of Ambainis et al. [5].

In many situations it makes sense to restrict ourselves to QES that have identical plaintext and ciphertext spaces; due to correctness, this is equivalent to unitarity.

Definition 3.3 (Unitary scheme) *A QES (τ_K, E, D) is called unitary if the encryption and decryption maps are controlled unitaries, i.e., if there exists $V = \sum_k U_A^{(k)} \otimes |k\rangle\langle k|_K$ such that $E(X) = VXV^\dagger$.*

It is straightforward to prove that, for unitary schemes, ITS is equivalent to the statement that the encryption maps $\{E_k\}$ form a unitary 1-design. Note that unitarity of E_k and correctness imply unitarity of D_k.

3.2 A new notion of non-malleability

Definition. We consider a scenario involving a user Alice and an adversary Mallory. The scenario begins with Mallory preparing a tripartite state ϱ_{ABR} over three registers: the plaintext A, the reference R, and the side-information B. The registers A and R are given to Alice, while Mallory keeps B. Alice then encrypts A into a ciphertext C and then transmits (or stores) it in the open. Mallory now applies an attack map

$$\Lambda : \mathcal{B}(\mathcal{H}_C \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_C \otimes \mathcal{H}_{\tilde{B}}) \,.$$

Mallory keeps the (transformed) side-information $\tilde{B}$ and returns C to Alice. Finally, Alice decrypts C back to A, and the scenario ends. We are now interested

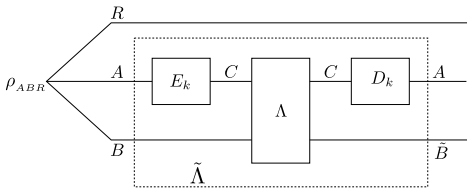

Fig. 1. The quantum non-malleability scenario.

in measuring the extent to which Mallory was able to increase her correlations with Alice's systems A and R. This can be understood by analyzing the mutual information $I(AR : \tilde{B})_{\tilde{\Lambda}(\varrho)}$ where $\tilde{\Lambda}_{AB\to A\tilde{B}}$ is the *effective channel* corresponding to Mallory's attack:

$$\tilde{\Lambda} = \mathrm{Tr}_K(D \circ \Lambda \circ E)((\cdot) \otimes \tau_K) \,. \tag{3.1}$$

We point out one way in which Mallory can always increase these correlations, regardless of the structure of the encryption scheme. First, she flips a coin b, and records the outcome in B. If $b = 1$, she replaces the contents of C with some fixed state σ_C, and otherwise she leaves C untouched. One then sees that Mallory's correlations have increased by $h(p_=(\Lambda, \varrho))$, where h denotes binary entropy and $p_=$ is a defined as follows.

$$p_=(\Lambda, \varrho) = \mathrm{Tr}\left[(\phi^+_{CC'} \otimes \mathbb{1}_{\tilde{B}})\Lambda(\phi^+_{CC'} \otimes \varrho_B)\right] . \tag{3.2}$$

This quantity is the inner product between the identity map and the map $\Lambda((\,\cdot\,)\otimes \varrho_B)$, expressed in terms of CJ states. Intuitively, it measures the probability with which Mallory chooses to apply the identity map; taking the binary entropy then gives us the information gain resulting from recording this choice.

We are now ready to define information-theoretic quantum non-malleability. Stated informally, a scheme is non-malleable if Mallory can only implement the attacks described above.

Definition 3.4 (Non-malleability) *A QES (τ_K, E, D) is non-malleable (NM) if for any state ϱ_{ABR} and any CPTP map $\Lambda_{CB\to C\tilde{B}}$, we have*

$$I(AR : \tilde{B})_{\tilde{\Lambda}(\varrho)} \leq I(AR : B)_\varrho + h(p_=(\Lambda, \varrho)). \tag{3.3}$$

One might justifiably wonder if the term $h(p_=(\Lambda, \varrho))$ is too generous to the adversary. However, as we showed above, every scheme is vulnerable to an attack which gains this amount of information. This term also appears (somewhat disguised) in the classical setting. In fact, if a classical encryption scheme satisfies Definition 3.4 against classical adversaries, then it also satisfies classical information-theoretic non-malleability as defined in [20].

Definition 3.4 directly generalizes the classical information-theoretic definition from [20]. In some settings, it might be preferable to have a definition which characterizes the set of effective attack channels as was done in [6]. As it turns out, NM can be defined in this way.

Theorem 3.7 (Non-malleability, alternative form) *A QES(τ, E, D) is NM if and only if for any attack $\Lambda_{CB\to C\tilde{B}}$, the effective map $\tilde{\Lambda}_{AB\to A\tilde{B}}$ has the form*

$$\tilde{\Lambda} = \mathrm{id}_A \otimes \Lambda'_{B\to\tilde{B}} + \frac{1}{|C|^2 - 1}\left(|C|^2 \langle D_K(\tau)\rangle - \mathrm{id}\right)_A \otimes \Lambda''_{B\to\tilde{B}} \tag{3.4}$$

where $\Lambda' = \mathrm{Tr}_{CC'}[\phi^+_{CC'}\Lambda(\phi^+_{CC'} \otimes (\cdot))]$ and $\Lambda'' = \mathrm{Tr}_{CC'}[\Pi^-_{CC'}\Lambda(\phi^+_{CC'} \otimes (\cdot))]$.

The proof of this theorem is postponed to the results section below (proof sketch) and the appendix.

Finally, as we will show in later sections, Definition 3.4 implies ABW-NM (see Definition 3.8), and schemes satisfying Definition 3.4 are sufficient for building quantum authentication under the strongest known definitions.

Non-malleability implies secrecy. In the classical case, non-malleability is independent from secrecy: the one-time pad is secret but malleable, and non-malleability is unaffected by appending the plaintext to each ciphertext. In the quantum case, on the other hand, we can show that NM implies secrecy. This is analogous to the fact that "quantum authentication implies encryption" [7].

Proposition 3.5 *Let (τ_K, E, D) be an NM QES. Then (τ_K, E, D) is ITS.*

Proof. Let B, ϱ_{AB}, and $\sigma_{CBK} = E(\varrho_{AB} \otimes \tau_K)$ be as in the definition of ITS (Definition 3.2). We first rename B to R. We then consider the non-malleability property in the following special-case scenario. The initial side-information register is empty, the final side-information register $\tilde{B}$ satisfies $\mathcal{H}_{\tilde{B}} \cong \mathcal{H}_C$, and the adversary map $\Lambda_{C \to C\tilde{B}}$ is defined as follows. Note that the "ciphertext-extraction" map $\Theta_{C \to C\tilde{B}} = \mathrm{id}_{C \to \tilde{B}}(\cdot) \otimes \tau_C$ has CJ state $\eta^{\Theta}_{CC'\tilde{B}} = \phi^{+}_{C'\tilde{B}} \otimes \tau_C$. We choose Λ so that its CJ state satisfies

$$\eta^{\Lambda}_{CC'\tilde{B}} = \frac{d^2}{d^2-1} \Pi^{-}_{CC'} \, \eta^{\Theta}_{CC'\tilde{B}} \, \Pi^{-}_{CC'} \, . \tag{3.5}$$

Applying the above projection to the CJ state of Θ ensures that Λ will have $p_{=}(\Lambda) = 0$ (note: $p_{=}(\Theta) > 0$.)

Direct calculation of the $C'\tilde{B}$ marginal of the CJ state of Λ yields

$$\eta^{\Lambda}_{C'\tilde{B}} = \frac{d^2-2}{d^2-1} \phi^{+}_{C'\tilde{B}} + \frac{1}{d^2-1} \tau_{C'} \otimes \tau_{\tilde{B}}. \tag{3.6}$$

This implies that the output $\sigma_{AR\tilde{B}} = \tilde{\Lambda}_{A \to A\tilde{B}}(\varrho_{AB})$ of the effective channel $\tilde{\Lambda}$ will satisfy

$$\sigma_{\tilde{B}R} = \frac{d^2-2}{d^2-1} \gamma_{\tilde{B}R} + \frac{1}{d^2-1} \tau_{\tilde{B}} \otimes \varrho_R, \tag{3.7}$$

where $\gamma_{CR} = (E_K)_{A \to C}(\varrho_{AR})$ and we used the fact that $\mathcal{H}_{\tilde{B}} \cong \mathcal{H}_C$. By non-malleability, we have

$$I(\tilde{B} : R)_\sigma + I(\tilde{B} : A|R)_\sigma = I(\tilde{B} : AR)_\sigma = 0. \tag{3.8}$$

In particular, $I(\tilde{B} : R)_\sigma = 0$ and thus $\sigma_{\tilde{B}R} = \sigma_{\tilde{B}} \otimes \varrho_R$. It follows by Equation (3.7) that

$$\gamma_{\tilde{B}R} = \frac{d^2-1}{d^2-2} \left(\sigma_{\tilde{B}} - \frac{1}{d^2-1} \tau_{\tilde{B}} \right) \otimes \varrho_R, \tag{3.9}$$

i.e., $\gamma_{\tilde{B}R}$ is a product state. This is precisely the definition of information-theoretic secrecy. □

Characterization of non-malleable schemes. Next, we provide a characterization of non-malleable schemes. First, we show that unitary schemes are equivalent to encryption with a unitary 2-design.

Theorem 3.6 *A unitary QES (τ_K, E, D) is NM if and only if $\{E_k\}_{k \in K}$ is a unitary 2-design.*

This fact is particularly intuitive when the 2-design is the Clifford group, a well-known exact 2-design. In that case, a Pauli operator acting on only one ciphertext qubit will be "propagated" (by the encryption circuit) to a completely random Pauli on all plaintext qubits. The plaintext is then maximally mixed, and the adversary gains no information. The Clifford group thus yields a perfectly non-malleable (and perfectly secret) encryption scheme using $O(n^2)$ bits of key [1].

It will be convenient to prove Theorem 3.6 as a consequence of our general characterization theorem, which is as follows.

Theorem 3.7 *Let (τ, E, D) be a QES. Then (τ, E, D) is NM if and only if, for any attack $\Lambda_{CB\to C\tilde{B}}$, the effective map $\tilde{\Lambda}_{AB\to A\tilde{B}}$ has the form*

$$\tilde{\Lambda} = \mathrm{id}_A \otimes \Lambda'_{B\to\tilde{B}} + \frac{1}{|C|^2-1}\left(|C|^2 \langle D_K(\tau)\rangle - \mathrm{id}\right)_A \otimes \Lambda''_{B\to\tilde{B}} \tag{3.10}$$

where $\Lambda' = \mathrm{Tr}_{CC'}[\phi^+_{CC'}\Lambda(\phi^+_{CC'}\otimes(\cdot))]$ and $\Lambda'' = \mathrm{Tr}_{CC'}[\Pi^-_{CC'}\Lambda(\phi^+_{CC'}\otimes(\cdot))]$.

We remark that the forward direction holds even if (τ, E, D) only fulfills the NM condition (Equation (3.3)) against adversaries with empty side-information B. The proof of Theorem 3.7 (with this strengthening) is sketched below. The full proof is somewhat technical and can be found in Appendix B. More precisely, we prove the stronger Theorem B.3, which implies the above by setting $\varepsilon = 0$.

Proof sketch. The first implication, i.e. NM implies Equation (3.10), is best proven in the Choi-Jamoiłkowski picture. Here, any QES defines a map

$$\mathcal{E}_{CC'\to AA'} = \frac{1}{|K|}\sum_k D_k \otimes E_k^T, \tag{3.11}$$

where the transpose E_k^T is the map whose Kraus operators are the transposes of the Kraus operators of E_k (in the standard basis). Our goal is to prove that this map essentially acts like the $U\bar{U}$-twirl. We decompose the space $\mathcal{H}_C^{\otimes 2}$ as

$$\mathcal{H}_C^{\otimes 2} = \mathbb{C}|\phi+\rangle \oplus \mathrm{supp}\Pi^- \tag{3.12}$$

which induces a decomposition of

$$\begin{aligned}\mathcal{B}(\mathcal{H}_C^{\otimes 2}) = \mathbb{C}|\phi^+\rangle\langle\phi^+| \oplus \left\{|\phi^+\rangle\langle v| \,\middle|\, \langle\phi^+ \mid v\rangle = 0\right\} \\ \oplus \left\{|v\rangle\langle\phi^+| \,\middle|\, \langle\phi^+ \mid v\rangle = 0\right\} \oplus \left\{X \in B \middle| \langle\phi^+|X = X|\phi^+\rangle = 0\right\}.\end{aligned} \tag{3.13}$$

On the first and last direct summands, the correct behavior of $\mathcal{E}$ is easy to show: the first one corresponds to the identity, and the last one to the non-identity channels Λ with $p_=(\Lambda) = 0$. For the remaining two spaces, we employ Lemma A.3 which shows that the encryption map of any valid encryption scheme has the form of appending an ancillary mixed state and then applying an isometry. Evaluating $\mathcal{E}(|\phi+\rangle\langle v|)$ for $\langle\phi^+ \mid v\rangle = 0$ reduces to evaluating the adjoint of the average encryption map, $E_K^\dagger$, on traceless matrices. It is, however, easy to verify that

$$\mathrm{Tr}_A \mathcal{E}_{CC'\to AA'}(\sigma_C \otimes (\cdot)_{C'}) = (E_K^T)_{C'\to A'}$$

for any σ_C. This can be used to prove $E_K = \langle \tau_C \rangle$ by observing that $\langle \phi^+|_{CC'} \sigma_C \otimes \varrho_{C'} |\phi^+\rangle_{CC'} = \mathrm{Tr}(\sigma_C \varrho_C)$, so for rank-deficient ϱ we can calculate $\mathcal{E}_{CC' \to AA'}(\sigma_C \otimes (\cdot)_{C'})$ using what we have already proven.

The other direction is proven by a simple application of Lemma A.2. □

The fact that NM is equivalent to 2-designs (for unitary schemes) is a straightforward consequence of the above.

Proof. (of Theorem 3.6) First, assume (τ_K, E, D) is a unitary NM QES with $E_k = U_k(\cdot)U_k^\dagger$. Then it has $|C| = |A|$, and $D_K(\tau_C) = \tau_A$, so the conclusion of Theorem 3.7 in this case (i.e., Equation (3.10)) is exactly the condition for $\{U_k\}$ to be an exact channel twirl design and therefore an exact 2-design. If (τ_K, E, D), on the other hand, is a unitary QES and $\{U_k\}$ is a 2-design, then Equation (3.10) holds and the scheme is therefore NM according to Theorem 3.7.

Relationship to ABW non-malleability. Ambainis, Bouda and Winter give a different definition of non-malleability, expressed in terms of the effective maps that an adversary can apply to the plaintext by acting on the ciphertext produced from encrypting with a random key [6]. According to their definition, a scheme is non-malleable if the adversary can only apply maps from a very restricted class *when averaging over the key, and without giving side information to the active adversary*. Let us recall their definition here.

First, given a QES (τ_K, E, D), we define the set $S := \{D_K(\sigma_C) \,|\, \sigma_C \in \mathcal{B}(\mathcal{H}_C)\}$ consisting of all valid average decryptions. We then define the class C_A^S of all "replacement channels". This is the set of CPTP maps belonging to the space

$$\mathrm{span}_{\mathbb{R}}\{\mathrm{id}_A, (X \mapsto \mathrm{Tr}(X)\sigma_A) : \sigma_A \in S\}\,. \tag{3.14}$$

We then make the following definition, which first appeared in [6].

Definition 3.8 (ABW non-malleability) *A QES (τ_K, E, D) is ABW-non-malleable (ABW-NM) if it is ITS, and for all channels $\Lambda_{C\to C}$, we have*

$$\mathrm{Tr}_K\left[D_{CK\to AK} \circ \Lambda_{C\to C} \circ E_{AK\to CK}(\,\cdot \otimes \tau_K)\right] \in C_A^S. \tag{3.15}$$

As indicated in [6], an approximate version of Equation (3.15) is obtained by considering the diamond-norm distance between the effective channel and the set C_A^S; this implies the possibility of an auxiliary reference system, which is denoted R in NM. We emphasize that this reference system is not under the control of the adversary. In particular, ABW-NM does not allow for adversaries which maintain *and actively use* side information about the plaintext system.

Another notable distinction is that [6] includes a secrecy assumption in the definition of an encryption scheme; under this assumption, it is shown that a unitary QES is ABW-NM if and only if the encryption unitaries form a 2-design. By our Theorem 3.6, we see that NM and ABW-NM are equivalent in the case of unitary schemes. So, in that case, ABW-NM actually ensures a much stronger security notion than originally considered by the authors of [6].

In the general case, NM is strictly stronger than ABW-NM. First, by comparing the conditions of Definition 3.8 to Equation (3.10), we immediately get the following corollary of Theorem 3.7.

Corollary 3.9 *If a QES satisfies NM, then it also satisfies ABW-NM.*

Second, we give a separation example which shows that ABW-NM is highly insecure; in fact, it allows the adversary to "inject" a plaintext of their choice into the ciphertext. This is insecure even under the classical definition of information-theoretic non-malleability of [20]. We now describe the scheme and this attack.

Example 3.10 *Suppose (τ_K, E, D) is a QES that is both NM and ABW-NM. Define a modified scheme (τ_K, E', D'), with enlarged ciphertext space $\mathcal{H}_{C'} = \mathcal{H}_C \oplus \mathcal{H}_{\hat{A}}$ (where $\mathcal{H}_{\hat{A}} \cong \mathcal{H}_A$) and encryption and decryption defined by*

$$\begin{aligned} E'(X) &= E(X)_C \oplus 0_{\hat{A}} \\ D'(X) &= D_{CK\to AK}(\Pi_C X \Pi_C) + \mathrm{id}_{\hat{A}K\to AK}(\Pi_{\hat{A}} X \Pi_{\hat{A}})\,. \end{aligned}$$

Then (τ_K, E', D') is ABW-NM but not NM.

While encryption ignores $\mathcal{H}_{\hat{A}}$, decryption measures if we are in C or $\hat{A}$ and then decrypts (in the first case) or just outputs the contents (in the second case.) This is a dramatic violation of NM: set $\mathcal{H}_{\tilde{B}} \cong \mathcal{H}_A$, trivial B and R, and

$$\Lambda_{C'\to C'\tilde{B}}(X) = \mathrm{Tr}(X) 0_C \oplus |\phi^+\rangle\langle\phi^+|_{\hat{A}\tilde{B}}\,; \tag{3.16}$$

it follows that, for all ϱ,

$$I(AR:\tilde{B})_{\tilde{\Lambda}(\varrho)} = 2\log|A| \gg h(|C'|^{-2}) = h(p_=(\Lambda,\varrho))\,. \tag{3.17}$$

Now let us show that (τ, E', D') is still ABW-NM. Let $\Lambda_{C'\to C'}$ be an attack, i.e., an arbitrary CPTP map. Then the effective plaintext map is

$$\tilde{\Lambda}_{A\to A} = D \circ \Lambda^C_{C\to C} \circ E + \Lambda^{\hat{A}}_{C\to A} \circ E, \tag{3.18}$$

where $\Lambda^C(X_C) = \Pi_C \Lambda(X_C \oplus 0_{\hat{A}})\Pi_C$ and $\Lambda^{\hat{A}}(X_C) = \mathrm{id}_{\hat{A}\to A}(\Pi_{\hat{A}}\Lambda(X_C \oplus 0_{\hat{A}})\Pi_{\hat{A}})$. Since (τ, E, D) is ITS (Theorem 3.5), there exists a fixed state ϱ^0_C such that $E_K(\varrho_A) = \varrho^0_C$ for all ϱ_A. Since (τ, E, D) is ABW-NM, we also know that

$$\mathrm{Tr}_K \circ D \circ \Lambda^C_{C\to C} \circ E = \tilde{\Lambda}_1 \in C^S_A\,,$$

with $S = \{D_K(\sigma_C)\,|\,\sigma_C \in \mathcal{B}(\mathcal{H}_C)\}$. We therefore get

$$\tilde{\Lambda}_{A\to A} = \tilde{\Lambda}_1 + \langle \Lambda^{\hat{A}}(\varrho^0_C)\rangle \in C^{S'}_A, \tag{3.19}$$

with $S' = \{D'_K(\sigma_{C'})\,|\,\sigma_{C'} \in \mathcal{B}(\mathcal{H}_{C'})\}$. This is true because S' contains all constant maps, as $D'_K(0_C \oplus \varrho_{\hat{A}}) = \varrho_A$.

4 The approximate setting

We now consider the case of approximate non-malleability. Approximate schemes are relevant for several reasons. First, an approximate scheme with negligible error can be more efficient than an exact one: the most efficient construction of an exact 2-design requires a quantum circuit of $O(n \log n \log\log n)$ gates [13], where approximate 2-designs can be achieved with linear-length circuits [14]. Second, in practice, absolutely perfect implementation of all quantum gates is too much to expect—even with error-correction. Third, when passing to authentication one must allow for errors, as it is always possible for the adversary to escape detection (with low probability) by guessing the secret key.

For all these reasons, it is important to understand what happens when the perfect secrecy and perfect non-malleability requirements are slightly relaxed. In this section, we show that our definitions and results are stable under such relaxations, and prove several additional results for quantum authentication. We begin with the approximate-case analogue of perfect secrecy.

Definition 4.1 (Approximate secrecy) *Fix $\varepsilon > 0$. A QES (τ_K, E, D) is ε-approximately secret (ϵ-ITS) if, for any $\mathcal{H}_B$ and any ϱ_{AB}, setting $\sigma_{CBK} = E(\varrho_{AB} \otimes \tau_K)$ implies $I(C:B)_\sigma \leq \varepsilon$.*

Analogously to the exact case, unitary schemes satisfying approximate secrecy are equivalent to approximate one-designs (see the full version of this article [3]).

4.1 Approximate non-malleability

Definition. We now define a natural approximate-case analogue of NM, i.e., Definition 3.4. Let us briefly recall the context. The malleability scenario is described by systems A, C, B and R (respectively, plaintext, ciphertext, side-information, and reference), an initial tripartite state ϱ_{ABR}, and an attack channel $\Lambda_{CB\to C\tilde{B}}$. Given this data, we have the effective channel $\tilde{\Lambda}_{AB\to A\tilde{B}}$ defined in Equation (3.1) and the "unavoidable attack" probability $p_=(\Lambda, \varrho)$ defined in Equation (3.2). The new definition now simply relaxes the requirement on the increase of the adversary's mutual information.

Definition 4.2 (Approximate non-malleability) *A QES (τ_K, E, D) is ε-non-malleable (ε-NM) if for any state ϱ_{ABR} and any CPTP map $\Lambda_{CB\to C\tilde{B}}$, we have*

$$I(AR:\tilde{B})_{\tilde{\Lambda}(\varrho)} \leq I(AR:B)_\varrho + h(p_=(\Lambda,\varrho)) + \varepsilon. \tag{4.1}$$

We record the approximate version of Proposition 3.5, i.e., non-malleability implies secrecy. The proof is a straightforward adaptation of the exact case.

Proposition 4.3 *Let (τ_K, E, D) be an ε-NM QES. Then (τ_K, E, D) is 2ε-ITS.*

Non-malleability with approximate designs. Continuing as before, we now generalize the characterization theorems of non-malleability (Theorem 3.7 and Theorem 3.6) to the approximate case.

Theorem 4.4 *Let (τ, E, D) be a QES with ciphertext dimension $|C| = 2^m$ and $r > 0$ a sufficiently large constant. Then the following holds:*

1. *If (τ, E, D) is 2^{-rm}-NM, then for any attack $\Lambda_{CB\to C\tilde{B}}$, the effective map $\tilde{\Lambda}_{AB\to A\tilde{B}}$ is $2^{-\Omega(m)}$-close (in diamond norm) to*

$$\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}} = \mathrm{id}_A \otimes \Lambda'_{B\to\tilde{B}} + \frac{1}{|C|^2 - 1}\left(|C|^2 \langle D_K(\tau)\rangle - \mathrm{id}\right)_A \otimes \Lambda''_{B\to\tilde{B}},$$

   with Λ', Λ'' as in Theorem 3.7.
2. *Suppose that $\log|R| = O(2^m)$, where R is the reference register in Definition 4.2. Then there exists a constant r, such that if every attack $\Lambda_{CB\to C\tilde{B}}$ results in an effective map that is 2^{-rm}-close to $\tilde{\Lambda}^{\text{exact}}$, then the scheme is $2^{-\Omega(m)}$-NM.*

This theorem is proven with explicit constants in Appendix B as Theorem B.3. The condition on R required for the second implication is necessary, as the relevant mutual information can at worst grow proportional to the logarithm of the dimension according to the Alicki-Fannes inequality. This is not a very strong requirement, as it should be relatively easy for the honest parties to put a bound on their total memory.

Next, we record the corollary which states that, for unitary schemes, approximate non-malleability is equivalent to encryption with an approximate 2-design. The proof proceeds as in the exact case, now starting from Theorem 4.4.

Theorem 4.5 *Let $\Pi = (\tau_K, E, D)$ be a unitary QES for n-qubit messages and $f : \mathbb{N} \to \mathbb{N}$ a function that grows at most exponential. Then there exists a constant $r > 0$ such that*

1. *If $\{E_k\}$ is a $\Omega(2^{-rn})$-approximate 2-design and $\log|R| \le f(n)$, then Π is $2^{-\Omega(n)}$-NM.*
2. *If Π is $\Omega(2^{-rn})$-NM, then $\{E_k\}_{k\in K}$ is a $2^{-\Omega(n)}$-approximate 2-design.*

Relationship to approximate ABW. Recall that, in Section 3.2, we discussed the relationship between our notion of exact non-malleability and that of Ambainis et al. [6] (i.e., ABW-NM.) As we now briefly outline, our conclusions carry over to the approximate case without any significant changes.

As described in Equation (3") of [6], one first relaxes the notion of ABW-NM appropriately by requiring that the containment (3.15) in Definition 3.8 holds up to ε error in the diamond-norm distance. In the unitary case, both definitions are equivalent to approximate 2-designs (by the results of [6], and our Theorem 4.5). In the case of general schemes, the plaintext injection attack described in Example 3.10 again shows that approximate ABW-NM is insufficient, and that approximate NM is strictly stronger.

4.2 Authentication

Definitions. Our definitions of authentication will be faithful to the original versions in [16, 18], with one slight modification. When decryption rejects, our encryption schemes (Definition 3.1) output $\perp$ in the plaintext space, rather than setting an auxiliary qubit to a "reject" state. These definitions are equivalent in the sense that one can always set an extra qubit to "reject" conditioned on the plaintext being $\perp$ (or vice-versa). Nonetheless, as we will see below, this mild change has some interesting consequences.

We begin with the definition of Dupuis, Nielsen and Salvail [16], which demands that the effective average channel of the attacker ignores the plaintext.

Definition 4.6 (DNS Authentication [16]) *A QES (τ_K, E, D) is called ε-DNS-authenticating if, for any CPTP-map $\Lambda_{CB\to CB'}$, there exists CP-maps $\Lambda^{\mathsf{acc}}_{B\to\tilde{B}}$ and $\Lambda^{\mathsf{rej}}_{B\to\tilde{B}}$ such that $\Lambda^{\mathsf{acc}}+\Lambda^{\mathsf{rej}}$ is*[1] *TP, and for all ϱ_{AB} we have*

$$\left\|\mathrm{Tr}_K D(\Lambda(E(\varrho_{AB}\otimes\tau_K))) - (\Lambda^{\mathsf{acc}}(\varrho_{AB}) + |\perp\rangle\langle\perp|\otimes\Lambda^{\mathsf{rej}}(\varrho_B))\right\|_1 \leq \varepsilon\,. \tag{4.2}$$

An alternative definition was recently given by Garg, Yuen and Zhandry [18]. It asks that, *conditioned on acceptance*, with high probability the effective channel is close to a channel which ignores the plaintext.

Definition 4.7 (GYZ Authentication [18]) *A QES (τ_K, E, D) is called ε-GYZ-authenticating if, for any CPTP-map $\Lambda_{CB\to CB'}$, there exists a CP-map $\Lambda^{\mathsf{acc}}_{B\to\tilde{B}}$ such that for all ϱ_{AB}*

$$\left\|\Pi_{\mathsf{acc}}\, D(\Lambda(E(\varrho_{AB}\otimes\tau_K)))\,\Pi_{\mathsf{acc}} - \Lambda^{\mathsf{acc}}(\varrho_{AB})\otimes\tau_K\right\|_1 \leq \varepsilon\,. \tag{4.3}$$

Here Π_{acc} is the acceptance projector, i.e. projection onto $\mathcal{H}_A$ in $\mathcal{H}_A\oplus\mathbb{C}|\perp\rangle$.

A peculiar aspect of the original definition in [18] is that it does not specify the outcome in case of rejection, and is thus stated in terms of trace non-increasing maps. Of course, all realistic quantum maps must be CPTP; this means that the designer of the encryption scheme must still declare what to do with the contents of the plaintext register after decryption. Our notion of decryption makes one such choice (i.e., output $\perp$) which seems natural.

GYZ authentication implies DNS authentication. A priori, the relationship between Definition 2.2 in [16] and Definition 8 in [18] is not completely clear. On one hand, the latter is stronger in the sense that it requires success with high probability (rather than simply on average.) On the other hand, the former makes the additional demand that the ciphertext is untouched even if we reject. As we now show, GYZ-authentication in fact implies DNS-authentication.

[1] Note that there is a typographic error in [16] and [11] at this point of the definition. In those papers, the two effective maps are asked to sum to the identity (instead of just a TP map), which is impossible for many obvious choices of Λ.

Theorem 4.8 *Let (τ, E, D) be ε-totally authenticating for sufficiently small ε. Then it is $O(\sqrt{\varepsilon})$-DNS authenticating.*

Proof. Let $\Lambda_{CB\to C\tilde{B}}$ be a CPTP map and $\varepsilon \leq 62^{-2}$. By Definition 4.7 there exists a CP map $\Lambda'_{B\to\tilde{B}}$ such that for all states ϱ_{AB},

$$\|\Pi_a D(\Lambda(E(\varrho_{AB}\otimes\tau_K)))\Pi_a - \Lambda'(\varrho_{AB}\otimes\tau_K)))\|_1 \leq \varepsilon\,. \tag{4.4}$$

Assume for simplicity that $D = M_\perp \circ D$, where $M_\perp$ measures the rejection symbol versus the rest. (otherwise we can define a new decryption map that way.) Define the CP maps

$$\Lambda^{(1)}_{AB\to\tilde{B}} = \mathrm{Tr}_A \Pi_a \tilde{\Lambda}(\cdot) \qquad \Lambda^{(2)}_{AB\to\tilde{B}} = \langle\perp|_A \tilde{\Lambda}(\cdot)|\perp\rangle_A$$
$$\Lambda''_{B\to\tilde{B}} = \mathrm{Tr}_C \Lambda(E_K(\tau_A)\otimes(\cdot)).$$

By Theorem 15 in [18] we have

$$|E_K(\varrho_{ABR}) - E_K(\tau_A)\otimes\varrho_{BR}\|_1 \leq 14\sqrt{\varepsilon}, \tag{4.5}$$

which implies that

$$\|\mathrm{Tr}_A \otimes \Lambda'' - \mathrm{Tr}_C \circ \Lambda \circ E_K\|_\diamond \leq \hat{\varepsilon} := 14\sqrt{\varepsilon}. \tag{4.6}$$

Note that

$$\begin{aligned}\mathrm{Tr}_C\circ\Lambda\circ E_K &= \mathrm{Tr}_{CK}\circ\Lambda\circ E((\cdot)\otimes\tau_K)\\ &= \mathrm{Tr}_{AK}\circ D\circ\Lambda\circ E((\cdot)\otimes\tau_K) = \mathrm{Tr}_A\circ\tilde{\Lambda}.\end{aligned} \tag{4.7}$$

On the other hand, we also have that, by Equation (4.4),

$$\left\|\mathrm{Tr}_A\circ\tilde{\Lambda} - \mathrm{Tr}_A\otimes\Lambda' - \Lambda^{(2)}\right\| \leq \left\|\mathrm{Tr}_A\left(\Pi_a\tilde{\Lambda}(\cdot)\right) - \Lambda'\right\|_\diamond \leq \varepsilon \tag{4.8}$$

Combining Equations (4.6), (4.7) and (4.8), we get

$$\left\|\Lambda^{(2)} - \mathrm{Tr}_A\otimes(\Lambda''-\Lambda')\right\|_\diamond \leq \varepsilon + \hat{\varepsilon}. \tag{4.9}$$

Now observe that

$$[\mathrm{Tr}_A\otimes(\Lambda'-\Lambda'')_{B\to\tilde{B}}]\circ\Xi_{A\to A} = \mathrm{Tr}_A\otimes(\Lambda'-\Lambda'')_{B\to\tilde{B}} \tag{4.10}$$

For all CPTP maps $\Xi_{A\to A}$. We define $\Lambda'''_{B\to\tilde{B}} = \Lambda^{(2)}(\tau_A\otimes(\cdot))$ and calculate

$$\begin{aligned}\left\|\Lambda^{(2)} - \mathrm{Tr}_A\otimes\Lambda'''\right\|_\diamond \leq{}& \left\|\Lambda^{(2)} - \mathrm{Tr}_A\otimes(\Lambda''-\Lambda')\right\|_\diamond\\ &+ \left\|\mathrm{Tr}_A\otimes(\Lambda''-\Lambda') - \mathrm{Tr}_A\otimes\Lambda'''\right\|_\diamond,\end{aligned}$$

by the triangle inequality for the diamond norm. Continuing with the calculation,

$$\begin{aligned}\left\|\Lambda^{(2)} - \mathrm{Tr}_A\otimes\Lambda'''\right\|_\diamond &\leq \varepsilon + \hat{\varepsilon} + \left\|\mathrm{Tr}_A\otimes(\Lambda''-\Lambda') - \mathrm{Tr}_A\otimes\Lambda'''\right\|_\diamond\\ &= \varepsilon + \hat{\varepsilon} + \left\|\left[\mathrm{Tr}_A\otimes(\Lambda''-\Lambda') - \Lambda^{(2)}\right]\circ\langle\tau_A\rangle_{A\to A}\right\|_\diamond\\ &\leq 2(\varepsilon+\hat{\varepsilon}) = 28\sqrt{\varepsilon} + 2\varepsilon.\end{aligned} \tag{4.11}$$

The first inequality above is Equation (4.9). The first equality is just a rewriting of the definition of Λ''', and the second equality is Equation (4.10). Finally, the last inequality is due to Equation (4.9) and the fact that the diamond norm is submultiplicative.

We have almost proven security according to Definition 4.6, as we have shown $\tilde{\Lambda}$ to be close in diamond norm to $\mathrm{id}_A \otimes \Lambda' + \langle|\perp\rangle\langle\perp|\rangle \otimes \Lambda'''$. However, $\Lambda' + \Lambda'''$ is only approximately TP; more precisely, we have that for all ϱ_{ABR},

$$|\mathrm{Tr}(\Lambda' + \Lambda''')(\varrho_{ABR}) - 1| \leq 28\sqrt{\varepsilon} + 3\varepsilon \tag{4.12}$$

by the triangle inequality. We therefore have to modify $\Lambda' + \Lambda''$ so that it becomes TP, while keeping the structure required for DNS authentication. Let $M_B = (\Lambda' + \Lambda''')^\dagger(\mathbb{1}_{\tilde{B}})$. (4.12). Defining the CP-map $\mathcal{M}(X) = M^{-1/2} X M^{-1/2}$ and noting it is well-behaved for small ε, it follows from a straightforward calculation (see the full version [3] of this article for details) that

$$\left\| \tilde{\Lambda}_{AB\to A\tilde{B}} - \mathrm{id}_A \otimes \Lambda^{\mathsf{acc}}_{B\to\tilde{B}} - \perp \otimes \Lambda^{\mathsf{rej}}_{B\to\tilde{B}} \right\|_\diamond \leq O(\sqrt{\varepsilon}). \tag{4.13}$$

with $\lambda^{\mathsf{acc}} = \Lambda' \circ \mathcal{M}$ and $\Lambda^{\mathsf{rej}} = \Lambda'' \circ \mathcal{M}$.

□

Achieving GYZ authentication with two-designs. In [18], the authors provide a scheme for their notion of authentication based on unitary eight-designs. We now show that, in fact, an approximate 2-design suffices. This implies that the well-known Clifford scheme (see e.g [15, 11]) satisfies the strong security of Definition 4.7. We remark that our proof is inspired by the reasoning based on Schur's lemma used in results on decoupling [8, 17, 24, 9].

Theorem 4.9 *Let* $\mathrm{D} = \{U_k\}_k$ *be a* δ*-approximate unitary 2-design on* $\mathcal{H}_C$*. Let* $\mathcal{H}_C = \mathcal{H}_A \otimes \mathcal{H}_T$ *and define*

$$E_k(X_A) = U_k \left(X_A \otimes |0\rangle\langle 0|_T\right) (U_k)^\dagger$$
$$D_k(Y_C) = \langle 0|_T (U_k)^\dagger Y U_k |0\rangle_T + \mathrm{Tr}((\mathbb{1}_T - |0\rangle\langle 0|_T)(U_k)^\dagger Y U_k)|\perp\rangle\langle\perp| .$$

Then the QES (τ_K, E, D) *is* $4(1/|T| + 3\delta)^{1/3}$*-GYZ-authenticating.*

Remark 4.10 *The following proof uses the same simulator as the proof for the 8-design scheme in [18], called "oblivious adversary" there. The construction exhibited there is efficient given that the real adversary is efficient.*

Proof. To improve readability, we will occasionally switch between adding subscripts to operators (indicating which spaces they act on) and omitting these subscripts. We begin by remarking that it is sufficient to prove the GYZ condition (specifically, Equation 4.3) for pure input states and isometric adversary channels. Indeed, for a general state ϱ_{AB} and a general map $\Lambda_{CB\to C\tilde{B}}$, we may

let ϱ_{ABR} and $V_{CB\to C\tilde{B}E}$ be the purification and Stinespring dilation, respectively. We then simply observe that the trace distance decreases under partial trace (see e.g. [25]). Let ϱ_{AB} be a pure input state and

$$\Lambda_{CB\to C\tilde{B}}(X_{CB}) = V_{CB\to C\tilde{B}} X_{CB} V_{CB\to C\tilde{B}}^\dagger$$

an isometry. We define the corresponding "ideal" channel Γ_V, and the corresponding "real, accept" channel Φ_k, as follows:

$$\begin{aligned}(\Gamma_V)_{B\to\tilde{B}} &= \frac{1}{|C|}\mathrm{Tr}_C V \text{ and} \\ (\Phi_k)_{AB\to A\tilde{B}} &= \langle 0|_T (U_k)_C^\dagger V_{CB\to C\tilde{B}} U_k |0\rangle_T. \end{aligned} \tag{4.14}$$

Note that for any matrix M with $\|M\|_\infty \leq 1$, the map $\Lambda_M(X) = M^\dagger X M$ is completely positive and trace non-increasing. We have

$$\|\Gamma_V\|_\infty \leq \frac{1}{|C|}\sum_i \|\langle i|V|i\rangle\|_\infty \leq 1. \tag{4.15}$$

We start by bounding the expectation of $\|((\Gamma_V)_{B\to\tilde{B}} - (\Phi_k)_{AB\to A\tilde{B}})|\varrho\rangle_{AB}\|_2^2$, as follows. To simplify notation, we set $\sigma_{ABT} := |\varrho\rangle\langle\varrho|_{AB} \otimes |0\rangle\langle 0|_T$ to be the tagged state corresponding to plaintext (and side information) ϱ_{AB}.

$$\begin{aligned}\frac{1}{|K|}\sum_k \|(\Gamma_V - \Phi_k)|\varrho\rangle\|_2^2 &= \frac{1}{|K|}\sum_k \langle\varrho|(\Gamma_V - \Phi_k)^\dagger(\Gamma_V - \Phi_k)|\varrho\rangle \\ &= \frac{1}{|K|}\sum_k \mathrm{Tr}\left[\sigma_{ABT}(U_k)^\dagger V^\dagger U_k |0\rangle\langle 0|(U_k)^\dagger V U_k\right] \\ &\quad - 2\frac{1}{|K|}\sum_k \mathrm{Tr}\left[\sigma_{ABT}(U_k)^\dagger V^\dagger U_k \Gamma_V\right] + \langle\varrho|\left(\Gamma_V\right)^\dagger \Gamma_V |\varrho\rangle\,. \end{aligned} \tag{4.16}$$

First we bound the second term, using the fact that Γ_V only acts on B.

$$\begin{aligned}\frac{1}{|K|}\sum_k \mathrm{Tr}\left[\sigma_{ABT}(U_k)^\dagger V^\dagger U_k \Gamma_V\right] &= \frac{1}{|K|}\sum_k \mathrm{Tr}\left[U_k \sigma_{ABT}(U_k)^\dagger V^\dagger \Gamma_V\right] \\ &= \int \mathrm{Tr}\left[\left(U\sigma_{ABT}U^\dagger + \Delta\right)V^\dagger\Gamma_V\right] \geq \int \mathrm{Tr}\left[U\sigma_{ABT}U^\dagger V^\dagger \Gamma_V\right] - \delta \\ &= \int \mathrm{Tr}\left[\sigma_{ABT}U^\dagger V^\dagger U \Gamma_V\right] - \delta = \langle\varrho|\left(\Gamma_V\right)^\dagger\Gamma_V|\varrho\rangle - \delta\,. \end{aligned} \tag{4.17}$$

In the above, the operator Δ is the "error" operator in the δ-approximate 2-design. The second equality above follows from $\|\Delta\|_1 \leq \delta$ and the fact that a 2-design is also a 1-design; the inequality follows by Hölder's inequality, and the last step follows from Schur's lemma.

The first term of the RHS of Equation (4.16) can be simplified as follows. We will begin by applying the swap trick (Lemma 2.1) $\mathrm{Tr}[XY] = \mathrm{Tr}[FX \otimes Y]$ in the

second line below. The swap trick is applied to register CC', with the operators X and Y defined as indicated below.

$$\begin{aligned}
&\frac{1}{|K|}\sum_k \mathrm{Tr}\Big[\underbrace{\sigma_{ABT}(U_k)^\dagger_C V^\dagger_{C\tilde{B}\to CB}(U_k)_C|0\rangle\langle 0|_T}_{X}\ \underbrace{(U_k)^\dagger_C V_{CB\to C\tilde{B}}(U_k)_C}_{Y}\Big] \\
&= \frac{1}{|K|}\sum_k \mathrm{Tr}\left[(\sigma_{ABT}\otimes|0\rangle\langle 0|_{T'})\left(U_k^{\otimes 2}\right)_{CC'} V^\dagger_{C\tilde{B}\to CB}V_{C'B\to C'\tilde{B}}\left(U_k^{\otimes 2}\right)^\dagger_{CC'} F_{CC'}\right] \\
&= \frac{1}{|K|}\sum_k \mathrm{Tr}\left[\left(U_k^{\otimes 2}\right)^\dagger_{CC'}(\sigma_{ABT}\otimes|0\rangle\langle 0|_{T'})\left(U_k^{\otimes 2}\right)_{CC'} V^\dagger_{C\tilde{B}\to CB}V_{C'B\to C'\tilde{B}}F_{CC'}\right] \\
&\le \int \mathrm{Tr}\left[\left(U^{\otimes 2}\right)^\dagger_{CC'}(\sigma_{ABT}\otimes|0\rangle\langle 0|_{T'})\,U^{\otimes 2}_{CC'}V^\dagger_{C\tilde{B}\to CB}V_{C'B\to C'\tilde{B}}F_{CC'}\right]+\delta \\
&= \int \mathrm{Tr}\left[(\sigma_{ABT}\otimes|0\rangle\langle 0|_{T'})\,U^{\otimes 2}_{CC'}V^\dagger_{C\tilde{B}\to CB}V_{C'B\to C'\tilde{B}}\left(U^{\otimes 2}\right)^\dagger_{CC'}F_{CC'}\right]+\delta.
\end{aligned} \tag{4.18}$$

The inequality above follows the same way as in Equation 4.17. Let $d = |C|$. An easy representation-theoretic calculation (see the Full version [3] for details) shows that

$$\int U^{\otimes 2}V^\dagger_{C\tilde{B}\to CB}V_{C'B\to C'\tilde{B}}\left(U^{\otimes 2}\right)^\dagger \mathrm{d}U = \mathbb{1}_{CC'}\otimes R^{\mathbb{1}}_B + F_{CC'}\otimes R^F_B, \tag{4.19}$$

where we have set

$$\begin{aligned}
R^{\mathbb{1}}_B &= \frac{1}{d(d^2-1)}\left(d^3\Gamma_V^\dagger\Gamma_V - d\mathbb{1}\right) = \frac{1}{(d^2-1)}\left(d^2\Gamma_V^\dagger\Gamma_V - \mathbb{1}\right) \\
R^F_B &= \frac{1}{d(d^2-1)}\left(d^2\mathbb{1} - d^2\Gamma_V^\dagger\Gamma_V\right) = \frac{d}{(d^2-1)}\left(\mathbb{1}-\Gamma_V^\dagger\Gamma_V\right).
\end{aligned}$$

plugging (4.19) into (4.18) and using Lemma 2.1 again, we get

$$\begin{aligned}
&\int \mathrm{Tr}\left[(\sigma_{ABT}\otimes|0\rangle\langle 0|_{T'})\,U^{\otimes 2}_{CC'}V^\dagger_{C\tilde{B}\to CB}V_{C'B\to C'\tilde{B}}\left(U^{\otimes 2}\right)^\dagger_{CC'}F_{CC'}\right] \\
&\quad= \mathrm{Tr}\left[(\sigma_{ABT}\otimes|0\rangle\langle 0|_{T'})\left(\mathbb{1}_{CC'}\otimes R^{\mathbb{1}}_{B^2\to\tilde{B}^2}+F_{CC'}\otimes R^F_{B^2\to\tilde{B}^2}\right)F_{CC'}\right] \\
&\quad= \mathrm{Tr}\left[|\varrho\rangle\langle\varrho|_B\left(R^{\mathbb{1}}_B+|A|R^F_B\right)\right] \\
&\quad= \mathrm{Tr}\left[|\varrho\rangle\langle\varrho|_B\left(\frac{d(d-|A|)}{d^2-1}\left(\Gamma_V^\dagger\Gamma_V\right)_B+\frac{d|A|-1}{d^2-1}\mathbb{1}_B\right)\right].
\end{aligned} \tag{4.20}$$

Now recall that $d = |A||T|$. Using the fact that $(a-1)/(b-1) \le a/b$ for $b \ge a$, we can give a bound as follows.

$$\begin{aligned}
&\mathrm{Tr}\left[|\varrho\rangle\langle\varrho|\left(\frac{d(d-|A|)}{d^2-1}\left(\Gamma_V^\dagger\Gamma_V\right)+\frac{d|A|-1}{d^2-1}\mathbb{1}\right)\right] \\
&\quad= \frac{d|A|(|T|-1)}{d^2-1}\langle\varrho|\left(\Gamma_V^\dagger\Gamma_V\right)|\varrho\rangle+\frac{d|A|-1}{d^2-1} \\
&\quad\le \langle\varrho|\left(\Gamma_V^\dagger\Gamma_V\right)|\varrho\rangle+\frac{1}{|T|}.
\end{aligned} \tag{4.21}$$

Putting everything together, we arrive at

$$\frac{1}{|K|}\sum_k \|(\Gamma_V - \Phi_k)|\varrho\rangle\|_2^2 \leq \frac{1}{|T|} + 3\delta. \tag{4.22}$$

By Markov's inequality this implies

$$\mathbb{P}\left[\left\|(\Gamma_V - \Phi_k)|\varrho\rangle\right\|_2^2 > \alpha\left(\frac{1}{|T|} + 3\delta\right)\right] \leq \frac{1}{\alpha} \tag{4.23}$$

which is equivalent to

$$\mathbb{P}\left[\left\|(\Gamma_V - \Phi_k)|\varrho\rangle\right\|_2 > \alpha^{1/2}\left(\frac{1}{|T|} + 3\delta\right)^{1/2}\right] \leq \frac{1}{\alpha}, \tag{4.24}$$

where the probability is taken over the uniform distribution on D. Choosing $\alpha = (1/|T| + 3\delta)^{-1/3}$ this yields

$$\mathbb{P}\left[\|(\Gamma_V - \Phi_k)|\varrho\rangle\|_2 > \left(\frac{1}{|T|} + 3\delta\right)^{1/3}\right] \leq \left(\frac{1}{|T|} + 3\delta\right)^{1/3}. \tag{4.25}$$

Let $S \subset D$ be such that $|S|/|\mathrm{D}| \geq 1 - (1/|T| + 3\delta)^{1/3}$ and $\|(\Gamma_V - \Phi_k)|\varrho\rangle\|_2 \leq (1/|T| + 3\delta)^{1/3}$ for all $U_k \in S$. Using the easy-to-verify inequality $\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_1 \leq 2\||\psi\rangle - |\phi\rangle\|_2$, we can bound

$$\begin{aligned}
&\frac{1}{|K|}\sum_{U_k\in\mathcal{D}} \left\|\Phi_k|\varrho\rangle\langle\varrho|\,(\Phi_k)^\dagger - \Gamma_V|\varrho\rangle\langle\varrho|\Gamma_V^\dagger\right\|_1 \\
&\quad\leq \frac{1}{|S|}\sum_{U_k\in\mathcal{S}} \left\|\Phi_k|\varrho\rangle\langle\varrho|\,(\Phi_k)^\dagger - \Gamma_V|\varrho\rangle\langle\varrho|\Gamma_V^\dagger\right\|_1 + 2\left(\frac{1}{|T|} + 3\delta\right)^{1/3} \\
&\quad\leq \frac{2}{|S|}\sum_{U_k\in\mathcal{S}} \|(\Gamma_V - \Phi_k)|\varrho\rangle\|_2 + 2|T|^{-1/3} \\
&\quad\leq 4\left(\frac{1}{|T|} + 3\delta\right)^{1/3}.
\end{aligned} \tag{4.26}$$

This completes the proof for pure states and isometric adversary channels. As noted above, the general case follows. □

As an example, one may set $|T| = 2^s$ (i.e. s tag qubits) and take an approximate unitary 2-design of accuracy 2^{-s}. The resulting scheme would then be $\Omega(2^{-s/3})$-GYZ-authenticating.

A straightforward corollary of the above result is that, in the case of unitary schemes, adding tags to non-malleable schemes results in GYZ authentication. We leave open the question of whether this is the case for general (not necessarily unitary) schemes.

Corollary 4.11 *Let (τ, E, D) be a 2^{-rn}-non-malleable unitary QES with plaintext space A. Define a new scheme (τ, E', D') with plaintext space A' where $A = TA'$ and*

$$E'(X) = E(X \otimes |0\rangle\langle 0|_T)$$
$$D'(Y) = \langle 0|_T D(Y)|0\rangle_T + \mathrm{Tr}\left[(\mathbb{1}_T - |0\rangle\langle 0|_T)D(Y)\right]|\perp\rangle\langle\perp|\,.$$

Then there is a constant $r > 0$ such that (τ, E', D') is $2^{-\Omega(n)}$-GYZ-authenticating if $|T| = 2^{\Omega(n)}$.

The proof is a direct application of Theorem 4.5 (approximate non-malleability is equivalent to approximate 2-design) and Theorem 4.9 (approximate 2-designs suffice for GYZ authentication.) We emphasize that, by Remark 2.3, exponential accuracy requirements can be met with polynomial-size circuits.

DNS authentication from non-malleability. We end with a theorem concerning the case of general (i.e., not necessarily unitary) schemes. We show that adding tags to a non-malleable scheme results in a DNS-authenticating scheme. In this proof we will denote the output system of the decryption map by $\overline{A}$ to emphasize that it is A enlarged by the reject symbol.

Theorem 4.12 *Let r be a sufficiently large constant, and let (τ, E, D) be an 2^{-rn}-NM QES with n qubit plaintext space A, and choose an integer d dividing $|A|$. Then there exists a decomposition $A = TA'$ and a state $|\psi\rangle_T$ such that $|T| = d$ and the scheme (τ, E', D') defined by*

$$E^t(X) = E(X \otimes |\psi\rangle\langle\psi|_T)$$
$$D^t(Y) = \langle\psi|_T D(Y)|\psi\rangle_T + \mathrm{Tr}\left[(\mathbb{1}_T - |\psi\rangle\langle\psi|_T)D(Y)\right]|\perp\rangle\langle\perp|\,.$$

is $(4/|T|) + 2^{-\Omega(n)}$-DNS-authenticating.

Proof. We prove the statement for $\varepsilon = 0$ for simplicity, the general case follows easily by employing Theorem 4.4 instead of Theorem 3.7.

By Theorem 3.7, for any attack map $\Lambda_{CB\to C\tilde{B}}$, the effective map is equal to

$$\tilde{\Lambda}_{AB\to\overline{A}\tilde{B}} = \mathrm{id}_A \otimes \Lambda'_{B\to\tilde{B}} + \frac{1}{|C|^2 - 1}\left(|C|^2\langle D_K(\tau_C)\rangle - \mathrm{id}\right)_{\overline{A}} \otimes \Lambda''_{B\to\tilde{B}} \tag{4.27}$$

for CP maps Λ' and Λ'' whose sum is TP. The effective map under the tagged scheme is therefore

$$\begin{aligned}\tilde{\Lambda}^t_{A'B\to\overline{A}'\tilde{B}} &= \langle\psi|_T \tilde{\Lambda}_{AB\to\overline{A}\tilde{B}}((\cdot)\otimes\psi_T)|\psi\rangle_T \\ &\quad + \mathrm{Tr}\big[(\mathbb{1}_T - \psi_T)\tilde{\Lambda}_{AB\to\overline{A}\tilde{B}}((\cdot)\otimes\psi_T)\big]|\perp\rangle\langle\perp| \\ &= (\mathrm{id}_{A'})_{A'\to\overline{A}'} \otimes \Lambda'_{B\to\tilde{B}} \\ &\quad + \left(|C|^2\left\langle\left(\langle\psi|_T D_K(\tau_C)|\psi\rangle_T\right)_{A'} \oplus \beta|\perp\rangle\langle\perp|\right\rangle - \mathrm{id}_{A'}\right)_{A\to\overline{A}'} \otimes \frac{\Lambda''_{B\to\tilde{B}}}{|C|^2 - 1}\end{aligned}$$

with $\beta = \mathrm{Tr}\left[(\mathbb{1} - \psi)_T D_K(\tau_C)\right]$. We would like to say that, unless the output is the reject symbol, the effective map on A is the identity. We do not know, however, what $D_K(\tau_C)$ looks like. Therefore we apply a standard reasoning that if a quantity is small *in expectation*, then there exists at least one small instance. We calculate the expectation of $\mathrm{Tr}\langle\psi|_T D_K(\tau_C)|\psi\rangle_T$ when the decomposition $A = TA'$ is drawn at random according to the Haar measure,

$$\begin{aligned}\int \mathrm{Tr}\langle\psi|U_A^\dagger D_K(\tau_C) U_A|\psi\rangle_T \mathrm{d}U_A &= \mathrm{Tr}\left[\left(\int U_A|\psi\rangle_T \otimes \mathbb{1}_{A'} \psi U_A^\dagger \mathrm{d}U_A\right) D_K(\tau_C)\right] \\ &= \frac{\mathrm{Tr}\mathbb{1}_A}{\mathrm{Tr}\Pi_{\mathsf{acc}}} \mathrm{Tr}\Pi_{\mathsf{acc}} D_K(\tau_C) \\ &\leq 1/|T|. \end{aligned} \tag{4.28}$$

Hence there exists at least one decomposition $A = TA'$ and a state $|\psi\rangle_T$ such that $\hat{\gamma} := \mathrm{Tr}\langle\psi|_T D_K(\tau_C)|\psi\rangle_T \leq 1/|T|$. Define $\gamma = \max(\hat{\gamma}, |C|^{-2})$. For the resulting primed scheme, let

$$\Lambda_{\mathsf{rej}} := \frac{(1-\gamma)|C|^2}{|C|^2 - 1}\Lambda'' \qquad \text{and} \qquad \Lambda_{\mathsf{acc}} = \Lambda' + \frac{\gamma|C|^2 - 1}{|C|^2 - 1}\Lambda''\,.$$

We calculate the diamond norm difference between the real effective map an the ideal effective map,

$$\begin{aligned}&\left\|\tilde{\Lambda}^t - \mathrm{id} \otimes \Lambda_{\mathsf{acc}} - \langle\perp\rangle\langle\perp| \otimes \Lambda_{\mathsf{rej}}\right\|_\diamond \\ &\leq \left\|\mathrm{id} \otimes \Lambda' + \frac{1}{|C|^2 - 1}\left(|C|^2\left\langle\left(\langle\psi|D_K(\tau)|\psi\rangle\right)\right\rangle - \mathrm{id}\right) \otimes \Lambda'' - \mathrm{id} \otimes \Lambda_{\mathsf{acc}}\right\|_\diamond \\ &\quad + \left\|\langle|\perp\rangle\rangle\langle\langle\perp|| \otimes (1-\hat{\gamma})|C|^2\Lambda''/(|C|^2 - 1) - \langle\perp\rangle\langle\perp| \otimes \Lambda_{\mathsf{rej}}\right\|_\diamond \\ &\leq (1 + |C|^{-2})(|T|^{-1} + 2|C|^{-2}) \\ &= |T|^{-1}(1 + (|A'||T|)^{-2})(1 + 2|A'|^{-2}) \\ &\leq 4|T|^{-1} \end{aligned} \tag{4.29}$$

as desired. □

Acknowledgments

The authors would like to thank Anne Broadbent, Alexander Müller-Hermes, Frédéric Dupuis and Christopher Portmann for helpful discussions. G.A. and C.M. acknowledge financial support from the European Research Council (ERC Grant Agreement 337603), the Danish Council for Independent Research (Sapere Aude) and VILLUM FONDEN via the QMATH Centre of Excellence (Grant 10059).

Bibliography

[1] Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Phys. Rev. A*, 70:052328, Nov 2004. doi: 10.1103/PhysRevA.70.052328.

[2] Dorit Aharonov, Michael Ben-Or, and Elad Eban. Interactive proofs for quantum computations. In *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*, pages 453–469, 2010.

[3] Gorjan Alagic and Christian Majenz. Quantum non-malleability and authentication. *CoRR*, abs/1610.04214, 2016. URL http://arxiv.org/abs/1610.04214.

[4] Robert Alicki and Mark Fannes. Continuity of quantum conditional information. *Journal of Physics A: Mathematical and General*, 37(5):L55, 2004.

[5] Andris Ambainis, Michele Mosca, Alain Tapp, and Ronald De Wolf. Private quantum channels. In *focs*, pages 547–553, 2000.

[6] Andris Ambainis, Jan Bouda, and Andreas Winter. Nonmalleable encryption of quantum information. *Journal of Mathematical Physics*, 50(4):042106, 2009.

[7] Howard Barnum, Claude Crépeau, Daniel Gottesman, Adam Smith, and Alain Tapp. Authentication of quantum messages. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 449–458. IEEE, 2002.

[8] Mario Berta, Matthias Christandl, and Renato Renner. The quantum reverse shannon theorem based on one-shot information theory. *Communications in Mathematical Physics*, 306(3):579–615, 2011.

[9] Mario Berta, Fernando GSL Brandao, Christian Majenz, and Mark M Wilde. Deconstruction and conditional erasure of quantum correlations. *arXiv preprint arXiv:1609.06994*, 2016.

[10] Fernando GSL Brandao, Aram W Harrow, and Michal Horodecki. Local random quantum circuits are approximate polynomial-designs. *arXiv preprint arXiv:1208.0692*, 2012.

[11] Anne Broadbent and Evelyn Wainewright. Efficient simulation for quantum message authentication. *arXiv preprint arXiv:1607.03075*, 2016.

[12] Man-Duen Choi. Completely positive linear maps on complex matrices. *Linear algebra and its applications*, 10(3):285–290, 1975.

[13] Richard Cleve, Debbie Leung, Li Liu, and Chunhao Wang. Near-linear constructions of exact unitary 2-designs. *Quantum Information and Computation*, 16(9&10):0721–0756, 2016.

[14] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Physical Review A*, 80(1):012304, 2009.

[15] Frédéric Dupuis, Jesper Buus Nielsen, and Louis Salvail. Secure two-party quantum evaluation of unitaries against specious adversaries. In *Annual Cryptology Conference*, pages 685–706. Springer, 2010.

[16] Frédéric Dupuis, Jesper Buus Nielsen, and Louis Salvail. Actively secure two-party evaluation of any quantum operation. In *Advances in Cryptology–CRYPTO 2012*, pages 794–811. Springer, 2012.

[17] Frédéric Dupuis, Mario Berta, Jürg Wullschleger, and Renato Renner. One-shot decoupling. *Communications in Mathematical Physics*, 328(1):251–284, 2014.

[18] Sumegha Garg, Henry Yuen, and Mark Zhandry. New security notions and feasibility results for authentication of quantum data. *arXiv preprint arXiv:1607.07759*, 2016.

[19] Andrzej Jamiołkowski. Linear transformations which preserve trace and positive semidefiniteness of operators. *Reports on Mathematical Physics*, 3 (4):275–278, 1972.

[20] Akinori Kawachi, Christopher Portmann, and Keisuke Tanaka. Characterization of the relations between information-theoretic non-malleability, secrecy, and authenticity. In *International Conference on Information Theoretic Security*, pages 6–24. Springer, 2011.

[21] Elliott H Lieb and Mary Beth Ruskai. A fundamental property of quantum-mechanical entropy. *Physical Review Letters*, 30(10):434, 1973.

[22] Elliott H. Lieb and Mary Beth Ruskai. Proof of the strong subadditivity of quantum-mechanical entropy. *Journal of Mathematical Physics*, 14(12): 1938–1941, 1973.

[23] Richard A Low. Pseudo-randomness and learning in quantum computation. *arXiv preprint arXiv:1006.5227*, 2010.

[24] Christian Majenz, Mario Berta, Frédéric Dupuis, Renato Renner, and Matthias Christandl. Catalytic decoupling of quantum information. *arXiv preprint arXiv:1605.00514*, 2016.

[25] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

[26] C. Portmann. Quantum authentication with key recycling. *ArXiv e-prints*, October 2016.

[27] W Forrest Stinespring. Positive functions on c*-algebras. *Proceedings of the American Mathematical Society*, 6(2):211–216, 1955.

A Technical lemmas

In the following we state some technical Lemmas that we need in this article. The proofs can be found in the full version [3].

Lemma A.1 *Let $X_{A\to B} \in L(\mathcal{H}_A, \mathcal{H}_B)$ be a linear operator from A to B. Then*

$$X_{A\to B}|\phi^+\rangle_{AA'} = \sqrt{\frac{|B|}{|A|}} X^T_{B'\to A'}|\phi^+\rangle_{BB'}. \tag{A.1}$$

The next group of lemmas is concerned with entropic quantities.

Lemma A.2 *Let $\Lambda^{(i)}_{A\to A'}$ be CPTP maps and $\Lambda^{(i)}_{B\to B'}$, $i=1,...,k$ CP maps for $i=1,...,k$ such that $\sum_i \Lambda^{(i)}_{B\to B'}$ is trace preserving. Let $\Lambda^{(i)}_{AB\to A'B'} = \Lambda^{(i)}_{A\to A'} \otimes \Lambda^{(i)}_{B\to B'}$ and define the CPTP maps*

$$\begin{aligned}\Lambda_{AB\to A'B'C} &= \sum_{i=1}^{k} \Lambda^{(i)}_{AB\to A'B'} \otimes |i\rangle\langle i|_C \text{ and}\\ \Lambda'_{B\to B'C} &= \sum_{i=1}^{k} \Lambda^{(i)}_{B\to B'} \otimes |i\rangle\langle i|_C.\end{aligned} \tag{A.2}$$

Then

$$I(A':B')_{\Lambda(\varrho)} \leq I(A:B)_\varrho + H(C|A)_{\Lambda'(\varrho)} \leq I(A:B)_\varrho + H(C)_{\Lambda(\varrho)} \tag{A.3}$$

for any quantum state ϱ_{AB}.

The final lemma characterizes CPTP maps that are invertible on their image such that the inverse is CPTP as well.

Lemma A.3 *Let (τ_K, E, D) be a QES. Then the encryption maps have the structure*

$$(E_k)_{A\to C} = (V_k)_{A\hat{C}\to C}\left((\cdot)\otimes\sigma^{(k)}_{\hat{C}}\right)(V_k)^\dagger_{A\hat{C}\to C}, \tag{A.4}$$

and the decryption maps hence must have the form

$$\begin{aligned}(D_k)_{C\to A} &= \mathrm{Tr}_{\hat{C}}\left[\Pi_{\mathrm{supp}\sigma^k}(V_k)^\dagger_{A\hat{C}\to C}(\cdot)(V_k)_{A\hat{C}\to C}\right]\\ &+ \left(\hat{D}_k\right)_{C\to A}\left[\left(\mathbb{1}_C - \Pi^{\mathrm{valid}}_k\right)(\cdot)\left(\mathbb{1}_C - \Pi^{\mathrm{valid}}_k\right)\right]\end{aligned} \tag{A.5}$$

for some quantum states $\sigma^{(k)}_{\hat{C}}$, isometries $(V_k)_{C\to A\hat{C}}$, and some CPTP map $\hat{D}_k$. Here, $\Pi^{\mathrm{valid}}_k = (V_k)_{A\hat{C}\to C}\,\Pi_{\mathrm{supp}\sigma^k}(V_k)^\dagger_{A\hat{C}\to C}$ is the projector onto the space of valid ciphertexts.

B Proof of characterization theorem

This section is dedicated to proving the characterization theorem for non-malleable QES, i.e., Theorem 4.4. We begin with two preparatory lemmas.

Lemma B.1 *For any QES (τ, E, D) the map $\mathcal{E} := |K|^{-1}\sum_k D_k \otimes E_k^T$ satisfies*

$$\mathcal{E}\left(|\phi^+\rangle\langle\phi^+|_{CC'}(X_C\otimes \mathrm{id}_{C'})\right) = \frac{|A|}{|C|}|\phi^+\rangle\langle\phi^+|_{AA'}(E_K^\dagger(X)\otimes \mathrm{id}_{A'})$$

This lemma is a consequence of the correctness condition and is proven in the full version [3] of this article.

Lemma B.2 *Suppose* (τ_K, E, D) *satisfies Definition 4.2 for trivial* B*. Then* $\mathcal{E} := |K|^{-1} \sum_k D_k \otimes E_k^T$ *satisfies*

$$\begin{aligned}
&\Bigg\| \mathcal{E}(X) - \frac{|A|}{|C|} \Bigg[\langle \phi^+ | X | \phi^+ \rangle | \phi^+ \rangle\langle \phi^+ | \\
&\quad + \operatorname{Tr}\left(\Pi^- X\right) \frac{1}{|C|^2 - 1} \left(|C|^2 D_K(\tau_C)_A \otimes \tau_{A'} - \phi^+_{AA'} \right) \Bigg] \Bigg\|_\diamond \\
&\leq 2\sqrt{2\varepsilon}|A| \left(2\sqrt{|A||C|} + 1 \right). \qquad \text{(B.1)}
\end{aligned}$$

Proof. It follows directly from the fact that (τ_K, E, D) is a QES together with Lemma A.1 that

$$\mathcal{E}(\phi^+_{CC'}) = \frac{|A|}{|C|} \phi^+_{AA'}. \qquad \text{(B.2)}$$

Let $\Lambda^{(i)}_{C \to C\tilde{B}_1}$, $i = 0, 1$ be two attack maps such that $\eta_{\Lambda^{(i)}} |\phi^+\rangle = 0$ for $i = 0, 1$ and define

$$\Lambda_{C \to C\tilde{B}_1\tilde{B}_2} = \frac{1}{2} \sum_{i=0,1} |i\rangle\langle i|_{\tilde{B}_2} \otimes \Lambda^{(i)}.$$

The the ε-NM property implies

$$I(AA' : \tilde{B}_1\tilde{B}_2)_{\eta_{\tilde{\Lambda}}} \leq \varepsilon,$$

and therefore, using Pinsker's inequality,

$$\begin{aligned}
&\Bigg\| \frac{1}{2} \sum_{i=0,1} |i\rangle\langle i|_{\tilde{B}} \otimes (\eta_{\tilde{\Lambda}^{(i)}})_{CC'\tilde{B}_1} \\
&- \frac{1}{4} \left(\sum_{i=0,1} |i\rangle\langle i|_{\tilde{B}} \otimes (\eta_{\tilde{\Lambda}^{(i)}})_{\tilde{B}_1} \right) \otimes \left(\sum_{i=0,1} (\eta_{\tilde{\Lambda}^{(i)}})_{CC'} \right) \Bigg\|_1 \leq \sqrt{2\varepsilon}. \qquad \text{(B.3)}
\end{aligned}$$

Observe that

$$\begin{aligned}
\eta_{\tilde{\Lambda}} &= \frac{1}{|K|} \sum_k D_k \circ \Lambda \circ E_k(\phi^+_{AA'}) \\
&= \frac{|C|}{|A|} \frac{1}{|K|} \sum_k \left(D_k \otimes E_k^T \right) \circ \Lambda(\phi^+_{CC'}) \\
&= \frac{|C|}{|A|} \mathcal{E} \circ \Lambda(\phi^+_{CC'}). \qquad \text{(B.4)}
\end{aligned}$$

Setting $(\eta_{\Lambda^{(0)}})_{CC'\tilde{B}_1} = \tau^-_{CC'} \otimes (\eta_{\Lambda^{(1)}})_{\tilde{B}_1}$, we get

$$\begin{aligned}\eta_{\tilde{\Lambda}^{(0)}} =& \frac{|C|}{|A|}\mathcal{E}(\tau^-) \otimes (\eta_{\Lambda^{(1)}})_{\tilde{B}_1} \\ =& \frac{|C|}{|A|}\frac{1}{|C|^2-1}\left(|C|^2\mathcal{E}(\tau_{CC'}) - \mathcal{E}(\phi^+_{CC'})\right) \otimes (\eta_{\Lambda^{(1)}})_{\tilde{B}_1} \\ =& \frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau_C)\otimes\tau_A - \phi^+_{AA'}\right) \otimes (\eta_{\Lambda^{(1)}})_{\tilde{B}_1}. \end{aligned} \tag{B.5}$$

and therefore

$$\begin{aligned}&\left\| \frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau_C)\otimes\tau_A - \phi^+_{AA'}\right) \otimes (\eta_{\Lambda^{(1)}})_{\tilde{B}_1} - \frac{|C|}{|A|}\mathcal{E}\left((\eta_{\Lambda^{(1)}})_{CC'\tilde{B}_1}\right)\right\|_1 \\ &\leq 2\sqrt{2\varepsilon}\end{aligned} \tag{B.6}$$

for all $\Lambda^{(1)}$. For any state $\varrho_{CC'\tilde{B}_1}$ with $\varrho_{CC'\tilde{B}}|\phi^+\rangle_{CC'} = 0$, we define the state

$$\begin{aligned}&\varrho'_{CC'\tilde{B}_1\tilde{B}_2}\frac{1}{C}\Big(|0\rangle\langle 0|_{\tilde{B}_2}\otimes\varrho_{CC'\tilde{B}_1} \\ &+ |1\rangle\langle 1|_{\tilde{B}_2}\otimes\left[\left((\mathbb{1}_C-\varrho_C)\otimes V_{C'}\right)\phi^+\left((\mathbb{1}_C-\varrho_C)\otimes V_{C'}\right)\right]\otimes\varrho_{\tilde{B}_2}\Big).\end{aligned} \tag{B.7}$$

Here, V is a unitary such that $\mathrm{Tr}(\mathbb{1}_C - \varrho_C)V_C^T = 0$. It is easy to see that such a unitary always exists, the existence is equivalent to the fact that any $|C|$-tuple of real numbers is the ordered list of side lengths of a polygon in the complex plain. Note that $\varrho'_{CC'\tilde{B}_1\tilde{B}_2}|\phi^+\rangle_{CC'} = 0$, and $\varrho'_{C'} = \tau_{C'}$. Together with the triangle inequality, equation (B.6) implies therefore that

$$\begin{aligned}&\frac{1}{|C|}\left\|\frac{|C|}{|A|}\mathcal{E}(\varrho) - \frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau_C)\otimes\tau_A - \phi^+_{AA'}\right)\otimes\varrho_{\tilde{B}_1}\right\|_1 \\ &+ \Bigg\| \frac{|C|}{|A|}\mathcal{E}\left[\left((\mathbb{1}_C-\varrho_C)\otimes V_{C'}\right)\phi^+\left((\mathbb{1}_C-\varrho_C)\otimes V_{C'}\right)\right] \\ &\quad -\frac{|C|-1}{|C|}\frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau_C)\otimes\tau_A - \phi^+_{AA'}\right)\Bigg\|_1 \leq 2\sqrt{2\varepsilon},\end{aligned}$$

i.e. in particular

$$\left\|\frac{|C|}{|A|}\mathcal{E}(\varrho) - \frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau_C)\otimes\tau_A - \phi^+_{AA'}\right)\otimes\varrho_{\tilde{B}_1}\right\|_1 \leq 2\sqrt{2\varepsilon}|C|.$$

As ϱ was arbitrary we have proven that

$$\left\|\frac{|C|}{|A|}\mathcal{E} - \left\langle\frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau_C)\otimes\tau_A - \phi^+_{AA'}\right)\right\rangle\right\|_\diamond \leq 2\sqrt{2\varepsilon}|C|. \tag{B.8}$$

The only fact that is left to show is, that $\|\mathcal{E}(|\phi^+\rangle\langle v|)\|_1$ is small for all normalized $|v\rangle$ such that $\langle\phi^+ \mid v\rangle = 0$. To this end, observe that $\mathrm{Tr}_A\circ\mathcal{E}(\sigma_C\otimes(\cdot)_{C'}) = E_K^T$

for all quantum states σ_C. Let ϱ_C be any quantum state that does not have full rank, note that such states span all of $\mathcal{B}(\mathcal{H}_C)$, and for hermitian operators there exists a decomposition into such operators that saturates the triangle inequality. Taking a quantum state σ_C such that $\langle\phi^+|\varrho\otimes\sigma|\phi^+\rangle = \frac{1}{|C|}\mathrm{Tr}\varrho_C\sigma_C^T = 0$ (the first equality is the mirror lemma A.1), we have

$$\left\|\mathcal{E}(\varrho\otimes\sigma) - \frac{|A|}{|C|}\frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau_C)\otimes\tau_A - \phi^+_{AA'}\right)\right\|_1 \leq 2\sqrt{2\varepsilon}|A|$$

according to what we have already proven. Using inequality (B.8) we arrive at

$$\left\|E_K^\dagger(X) - \frac{|A|}{|C|}\tau_A\mathrm{Tr}(X)\right\|_1 \leq 2\sqrt{2\varepsilon}|A|\|X\|_1 \tag{B.9}$$

For Hermitian matrices X and therefore

$$\left\|E_K^\dagger(X) - \frac{|A|}{|C|}\tau_A\mathrm{Tr}(X)\right\|_1 \leq 4\sqrt{2\varepsilon}|A|\|X\|_1 \tag{B.10}$$

For arbitrary X. We can write $|v\rangle_{CC'} = X_C|\phi^+\rangle_{CC'}$ for some traceless matrix X_C. Now we calculate

$$\begin{aligned}
\left\|\mathcal{E}(|\phi^+\rangle\langle v|_{CC'})\right\|_1 =& \left\|\frac{|A|}{|C|}|\phi^+\rangle\langle\phi^+|_{AA'}\left(E_K^\dagger(X^\dagger)\right)_A\right\|_1 \\
=& \frac{|A|}{|C|}\left\|\left(E_K^\dagger(X)\right)_A|\phi^+\rangle_{AA'}\right\|_2 \\
=& \frac{\sqrt{|A|}}{|C|}\left\|E_K^\dagger(X)\right\|_2 \\
\leq& \frac{\sqrt{|A|}}{|C|}\left\|E_K^\dagger(X)\right\|_1 \\
\leq& \frac{|A|^{3/2}}{|C|}4\sqrt{2\varepsilon}\|X\|_1 \\
\leq& 4\sqrt{2\varepsilon}|A|^{3/2}.
\end{aligned} \tag{B.11}$$

The first equation is Lemma B.1, the sencond and third equations are easily verified, the first inequality is a standard norm inequality, the second inequality is Equation (B.10), and the last inequality follows from the normalization of $|v\rangle$. By the Schmidt decomposition, we get a stabilized version of this inequality,

$$\left\|\mathcal{E}(|\phi^+\rangle_{CC'}|\alpha\rangle_{\tilde{B}_1}\langle v|_{CC'\tilde{B}_1})\right\|_1 \leq 2\sqrt{2\varepsilon}|A|^{3/2}, \tag{B.12}$$

for all $|\alpha\rangle_{\tilde{B}_1}$ and all $|v\rangle_{CC'\tilde{B}}$ such that $\langle\phi^+ \mid v\rangle = 0$ Combining everything we arrive at

$$\begin{aligned}
&\Big\| \mathcal{E}(X) - \frac{|A|}{|C|}\Big[\langle\phi^+|X|\phi^+\rangle|\phi^+\rangle\langle\phi^+| \\
&\quad + \operatorname{Tr}\left(\Pi^- X\right)\frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau_C)_A \otimes \tau_{A'} - \phi^+_{AA'}\right)\Big]\Big\|_\diamond \\
&\leq 2\sqrt{2\varepsilon}|A|\left(4\sqrt{|A|}+1\right). \qquad \text{(B.13)}
\end{aligned}$$

□

We are now ready to prove the characterization theorem Theorem 4.4 in the ε-approximate setting (including the exact case, Theorem 3.7 by setting $\varepsilon = 0$.)

Theorem B.3 (Precise version of Theorem 4.4) *Let $\Pi = (\tau, E, D)$ be a QES.*

1. *If Π is ε-NM, then any attack map $\Lambda_{CB\to C\tilde{B}}$ results in an effective map $\tilde{\Lambda}_{AB\to A\tilde{B}}$ fulfilling*

$$\left\|\tilde{\Lambda}_{AB\to A\tilde{B}} - \tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}\right\|_\diamond \leq 2\sqrt{2\varepsilon}|A|^4|C|\left(4\sqrt{|A|}+1\right), \qquad \text{(B.14)}$$

   where

$$\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}} = \mathrm{id}_A \otimes \Lambda'_{B\to\tilde{B}} + \frac{1}{|C|^2-1}\left(|C|^2\langle D_K(\tau)\rangle - \mathrm{id}\right)_A \otimes \Lambda''_{B\to\tilde{B}},$$

   with $\Lambda' = \operatorname{Tr}_{CC'}[\phi^+_{CC'}\Lambda(\phi^+_{CC'}\otimes(\cdot))]$ and $\Lambda'' = \operatorname{Tr}_{CC'}[\Pi^-_{CC'}\Lambda(\phi^+_{CC'}\otimes(\cdot))]$.
2. *Conversely, if for a scheme all effective maps fulfil Equation (B.14) with the right hand side replaced by ε, then it is $5\varepsilon(\log(|A|)+r)+3h(\varepsilon)$-NM, where r is a bound on the size of the honest user's side information.*

Proof. We start with *1.* We want to bound the diamond norm distance between the effective map $\tilde{\Lambda}$ resulting from an attack Λ and the idealized effective map $\tilde{\Lambda}^{\text{exact}}$. Let

$$|\psi\rangle_{AA'BB'} = \sum_{i=0}^{|A|^2-1}\sqrt{p_i}|\alpha_i\rangle_{AA'}\otimes|\beta_i\rangle_{BB'}$$

be an arbitrary pure state given in its Schmidt decomposition across the bipartition AA' vs. BB'. We can Write $|\alpha_i\rangle_{AA'} = X^{(i)}_{A'}|\phi^+\rangle$ for some matrices $X^{(i)}$ satisfying $\|X^{(i)}\|_\infty \leq |A|$. We calculate the action of $\tilde{\Lambda}$ on $|\alpha_i\rangle\langle\alpha_j|_{AA'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'}$,

$$\begin{aligned}
&\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}(|\alpha_i\rangle\langle\alpha_j|_{AA'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'}) = X^{(i)}_{A'}\Big(|\phi^+\rangle\langle\phi^+|_{AA'}\otimes\Lambda'_{B\to\tilde{B}}(|\beta_i\rangle\langle\beta_j|_{BB'}) \\
&+\frac{1}{|C|^2-1}\left(|C|^2 D_K(\tau)_A\otimes\tau_{A'} - |\phi^+\rangle\langle\phi^+|_{AA'}\right)\otimes\Lambda''_{B\to\tilde{B}}(|\beta_i\rangle\langle\beta_j|_{BB'})\Big)X^{(j)}_{A'}. \qquad \text{(B.15)}
\end{aligned}$$

In a similar way we get

$$\begin{aligned}
&\tilde{\Lambda}_{AB\to A\tilde{B}}(|\alpha_i\rangle\langle\alpha_j|_{AA'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'})\\
&=X^{(i)}_{A'}\tilde{\Lambda}_{AB\to A\tilde{B}}(|\phi^+\rangle\langle\phi^+|_{AA'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'})X^{(i)}_{A'}\\
&=\frac{|C|}{|A|}X^{(i)}_{A'}\mathcal{E}_{CC'\to AA'}\circ\Lambda_{CB\to C\tilde{B}}(|\phi^+\rangle\langle\phi^+|_{CC'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'})X^{(i)}_{A'}. \qquad \text{(B.16)}
\end{aligned}$$

Using Lemma B.2 we bound

$$\begin{aligned}
&\left\|\left(\tilde{\Lambda}_{AB\to A\tilde{B}}-\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}\right)(|\alpha_i\rangle\langle\alpha_j|_{AA'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'})\right\|_1\\
=&\left\|X^{(i)}_{A'}\left(\tilde{\Lambda}_{AB\to A\tilde{B}}-\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}\right)(|\phi^+\rangle\langle\phi^+|_{AA'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'})X^{(j)}\right\|_1\\
\le&\left\|X^{(i)}\right\|_\infty\left\|X^{(j)}\right\|_\infty\left\|\left(\tilde{\Lambda}_{AB\to A\tilde{B}}-\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}\right)(|\phi^+\rangle\langle\phi^+|_{AA'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'})\right\|_1\\
=&\left\|X^{(i)}\right\|_\infty\left\|X^{(j)}\right\|_\infty\Bigg\|\frac{|C|}{|A|}\mathcal{E}_{CC'\to AA'}\circ\Lambda_{CB\to C\tilde{B}}(|\phi^+\rangle\langle\phi^+|_{CC'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'})\\
&\qquad\qquad-\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}(|\phi^+\rangle\langle\phi^+|_{AA'}\otimes|\beta_i\rangle\langle\beta_j|_{BB'})\Bigg\|_1\\
\le&2\sqrt{2\varepsilon}|A|^2|C|\left(4\sqrt{|A|}+1\right). \qquad \text{(B.17)}
\end{aligned}$$

The inequalities result from applying Hölder's inequality twice, and Lemma B.2, respectively. Using the triangle inequality we get

$$\begin{aligned}
&\left\|\left(\tilde{\Lambda}_{AB\to A\tilde{B}}-\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}\right)(|\psi\rangle\langle\psi|_{AA'BB'})\right\|_1\\
\le&\,2\sqrt{2\varepsilon}|A|^2|C|\left(4\sqrt{|A|}+1\right)\sum_{i,j=0}^{|A|^2-1}\sqrt{p_ip_j}\\
\le&\,2\sqrt{2\varepsilon}|A|^4|C|\left(4\sqrt{|A|}+1\right). \qquad \text{(B.18)}
\end{aligned}$$

As $|\psi\rangle$ was arbitrary, we have proven

$$\left\|\tilde{\Lambda}_{AB\to A\tilde{B}}-\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}\right\|_\diamond\le2\sqrt{2\varepsilon}|A|^4|C|\left(4\sqrt{|A|}+1\right). \qquad \text{(B.19)}$$

Now let us prove 2. Let $\Lambda_{CB\to C\tilde{B}}$ again be an arbitrary attack map, and assume that the resulting effective map is ε-close to $\tilde{\Lambda}^{\text{exact}}_{AB\to A\tilde{B}}$. Observe that $p^{=}(\Lambda,\varrho)=\text{Tr}\Lambda'(\varrho_B)$.

By the Alicki-Fannes inequality [4] and Lemma A.2, this implies

$$I(AR:\tilde{B})_{\tilde{\Lambda}(\varrho)}\le I(AR:B)_\varrho+h(p^{=}(\Lambda,\varrho))+5\varepsilon\log(|A||R|)+3h(\varepsilon) \qquad \text{(B.20)}$$

with the help of Lemma A.2. □