Quantum steganography over noiseless channels: Achievability and bounds

Chris Sutherland[1] and Todd A. Brun[1,2]
[1]*Department of Physics, University of Southern California, Los Angeles, California 90089, USA*
[2]*Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089, USA*

(Received 16 May 2019; revised manuscript received 13 January 2020; accepted 16 March 2020; published 12 May 2020)

Quantum steganography is the study of hiding secret quantum information by encoding it into what an eavesdropper would perceive as an innocent-looking message. Here we study an explicit steganographic encoding for Alice to hide her secret message in the syndromes of an error-correcting code, so that the encoding simulates a given noisy quantum channel. We calculate achievable rates of steganographic communication over noiseless quantum channels using this encoding. We give definitions of secrecy and reliability for the communication process, and with these assumptions derive upper bounds on the amount of steganographic communication possible and show that these bounds match the communication rates achieved with our encoding. This gives a steganographic capacity for a noiseless channel emulating a given noisy channel.

DOI: 10.1103/PhysRevA.101.052319

I. INTRODUCTION

The study of steganography is perhaps best motivated by considering an example. Suppose two political protestors Alice and Bob are arrested and put into two widely separated jail cells. The warden allows them to communicate with handwritten letters that he reads before delivering. However, if the warden reads anything in the letters that he finds suspicious (such as a possible escape plan), then he will not deliver the letter. Luckily, Alice and Bob exchanged a secret key before their incarceration. Can Alice and Bob communicate their escape plan to each other without arousing the warden's suspicions? This is where the study of steganography comes into play.

The science of sending information through seemingly innocuous messages has a long history, dating back to at least 440 B.C. [1]. It is worth making clear its differences from cryptography. In cryptography, a secret message (the *plaintext*) is encrypted using the shared secret key and the resulting *ciphertext* is then sent to the desired receiver to be decoded. If an eavesdropper (Eve) observes the ciphertext, she cannot decode it without the secret key. However, she will know that there *is* a secret message since Alice is sending apparent gibberish to Bob.

By contrast, if Alice uses a steganographic encoding, she hides the secret message (or *stegotext*) into a larger *covertext*, which appears to Eve as an innocuous message. The hidden message may or may not be encrypted itself, but the main line of defense is that the eavesdropper is unaware that a message is even being sent.

During World War II, a Japanese spy named Velvalee Dickinson sent classified information to neutral South America. She was a dealer in dolls and her letters discussed the quantity and type of doll to ship. The covertext was the doll orders, while the concealed stegotext was encoded information about battleship movements [2].

The quantum analog of cryptography has been widely studied [3]. However, the quantum analog of steganography is still in a relatively early stage. There have been a number of different proposals for encoding quantum information steganographically, or encoding classical information into quantum states or channels [4,5]. In this paper, we consider hiding secret messages as error syndromes of a quantum error-correcting code [6]. This approach to quantum steganography has been studied in detail by Shaw and Brun, with explicit encoding and decoding procedures and calculated rates of communication and secret key consumption [7,8]. It was shown that such schemes can hide both quantum and classical information, with a quantitative measure of secrecy, even in the presence of a noisy physical channel. When the error rate of the physical channel is lower than the eavesdropper's expectation, it is possible to achieve nonzero asymptotic rates of communication. (If the eavesdropper has exact knowledge of the channel, secret communication may still be possible, but the amount of secret information that can be transmitted in general grows sublinearly with the number of channel uses.)

More recently, a closely related idea has been studied under the name of quantum covert communication [9–13]. Many of the ideas in this paper are closely related to steganographic requirements, such as secrecy and recoverability. This is not surprising since covert quantum communication can be seen as a special case of quantum steganography over noisy quantum channels in the case when the eavesdropper has exact knowledge of the channel, and where Eve assumes the channel is idle (so only noise is being transmitted). Similarly, quantum steganography is a type of covert quantum communication where Eve knows about the covertext communication but not the hidden stegotext, and where Eve may not have perfect knowledge of the channel. The work on covert communication has generally found that if Eve has exact knowledge of the channel, the amount of secret communication that can be done grows like the square root of the number of channel used.

There is also a somewhat related idea known as entropic security [14,15]. The sender chooses a message from a known message space and encrypts the message. For entropic

security, whenever an adversary intercepts an encryption, they should not be able to predict any function on the message, as long as the original message has sufficiently high entropy. This differs from steganography, however, in that Alice and Bob are using a shared secret key to avoid making the adversary suspicious at all. Should the adversary become suspicious, then could indeed intercept and read the message, unless the message is also encrypted using more shared secret key.

The goal of this paper is to formalize the assumptions and reasonable conditions of quantum steganography introduced in [7], and to give upper bounds on the achievable rates of quantum communication while remaining secure from an eavesdropper's suspicion, for the special case when the true underlying channel is noiseless. Our results include achievability results as well as converse proofs for quantum steganography.

In Sec. II, we formalize our notion of quantum steganography where secret messages are hidden in the syndromes of an error-correcting code and outline a specific steganographic encoding where Alice is able to emulate any general quantum channel $\mathcal{N}$ on her encoded secret message and covertext. We work out specific examples for the bit-flip channel and the depolarizing channel, before giving the more general result. In Sec. III, we prove upper bounds on the amount of steganographic communication possible and show that these bounds are asymptotically equal to the rates achieved in the previous section.

The assumption that the physical channel is noiseless greatly simplifies the analysis. However, we believe that the main intuition underlying this approach will apply equally well in the case of a noisy channel. We will end this paper with a discussion of how to extend this work to the case where the physical channel between the two parties is noisy.

II. QUANTUM STEGANOGRAPHY: ACHIEVABILITY

As discussed in Sec. I, there have been several approaches to generalizing steganography to the quantum setting [4–6]. Here we will make explicit the notion of quantum steganography based on syndromes of quantum error-correcting codes. We assume that Eve expects to see quantum information passing through a noisy quantum channel. However, the actual physical channel is assumed to be noiseless. This is obviously an idealized assumption, which greatly simplifies the analysis; we will discuss below how it might be justified at least as an approximation.

Alice wants to send a secret message steganographically to Bob. Using her shared secret key, she encodes the stegotext into a codeword of a quantum error-correcting code (QECC) with errors applied to it [6], and sends it to Bob. The codeword encodes an innocent state; the stegotext is conveyed in the errors. If Eve were to perform measurements on this codeword, it would be indistinguishable from an innocent encoded covertext that had passed through a given noisy quantum channel to Bob [7,8].

Before discussing how to quantify the security of a quantum steganographic protocol, let us make clear what Alice is trying to achieve. Alice wants to encode an innocent covertext state, together with her secret message, into an N-qubit codeword in such a way that it cannot be distinguished from the covertext alone encoded into a quantum error-correcting code that has undergone typical errors induced by the quantum channel $\mathcal{N}^{\otimes N}$. The steganographic encoding works by mapping all possible secret messages onto syndromes of the QECC. This encoding is not limited to classical messages: it is possible to encode a quantum state by preparing the codeword in a superposition of different error syndromes.

In analyzing this quantum steganography protocol, we make the following assumptions. Alice is communicating with Bob by a quantum channel that is actually noiseless. But the eavesdropper, Eve, believes that this channel is noisy, perhaps because Alice and Bob have been systematically making the channel appear noisier than it actually is. Because Alice and Bob have been systematically deceiving Eve in this way, we assume that they know (at least fairly closely) what Eve's estimate of the channel is. Before the protocol began, Alice and Bob shared with each other a secret key: an arbitrarily long string of random bits. This key is known only to the two of them. But once the protocol begins, they cannot communicate except through channels that can be monitored by Eve. Alice sends an innocent-looking message to Bob over the channel. This is a covertext state ρ_c, encoded into an error-correcting code; it is assumed that the choice of code is known to Eve, and this code should be a plausible choice for the noisy channel that Eve believes exists.

One important caveat for this section: we will be considering the case where the QECC that Alice uses is nondegenerate. That is, each typical error corresponds to a unique error syndrome. This allows Alice to communicate as much steganographic information as possible, and it allows us to ignore the details of which QECC is being used. Methods similar to those in this section should also work for degenerate codes; but in that case, the encoding will be strongly dependent on the properties of the particular code since the typical errors must first be grouped into equivalent sets and then the possible messages mapped onto these sets. We also use this assumption in the next section to get specific expressions for the upper bound on the secret communication rate.

To clarify how the encoding works, we start by considering two examples for relatively simple channels: first, the case where Alice is emulating a bit-flip channel $\mathcal{N}_p^{BF}$ on the codeword, and second, the case where she is emulating the depolarizing channel. Finally, we consider a more general error map $\mathcal{N}^{\otimes N}$. The message qubits are encoded into the error syndromes of the codeword of the QECC that she is using.

A. The bit-flip channel

Suppose that Eve believes the channel connecting Alice and Bob to be a bit-flip channel, with a probability p of error per qubit sent. (The actual physical channel is noiseless, as assumed above.) Alice sends a codeword of length N to Bob, encoding some "innocent" covertext state ρ_c. The errors in the codewords that Alice sends to Bob should be binomially distributed: pN is the mean number of errors of this distribution and the variance is $(1-p)pN$. The total probability that there is an error of weight w on the codeword should be

$$p_w = \binom{N}{w} p^w (1-p)^{N-w}. \tag{1}$$

There are

$$\binom{N}{w} \equiv \frac{N!}{w!(N-w)!}$$

such errors, all with equal probability $p^w(1-p)^{N-w}$.

If N is large, then it is extremely likely that the number of bit flips will be a *typical* error—that is, an error of weight w within a narrow range about the mean pN. Alice's encoding will make use of these typical errors. For each w from $Np(1-\delta)$ to $Np(1+\delta)$, where $\sqrt{(1-p)/pN} \ll \delta \ll 1$, Alice chooses at random a set of C_w possible error strings of weight w. (An *error string* of weight w is a string of N bits, with a 1 at every location with a bit flip and 0 at every location with no error.) This random choice is made using the shared secret key with Bob, so that Bob also knows which set of errors is being used to encode secret messages, but Eve (who does not share the key) could not know this.

Let these sets of error strings of weight w be called $\{S_w\}$, and the set of all strings used in the encoding is

$$S = \bigcup_w S_w. \tag{2}$$

We sum up

$$C = \sum_{w=Np(1-\delta)}^{Np(1+\delta)} C_w = |S|. \tag{3}$$

So the total number of strings in the set S is C. This number C is the total number of possible distinct secret messages that Alice can send to Bob (though she may also send *superpositions* of these messages). So the message encodes $M = \log_2 C$ bits (or qubits) of information.

Note that we are assuming all the messages to be equally likely. If the messages are not equally distributed, the messages can be first mapped to preliminary codewords that are equally distributed. The algorithm to do this would use more key, but still keep the dependance sublinear. Of course, we could properly encrypt the message before sending, but this would certainly increase the key rate from sublinear to linear.

Define the probability $q = 1/C$. These error strings S are typical strings (using the definition of weak typicality from information theory). Eve should not be suspicious at seeing such an error string since it matches a probable result for the channel that she expects. For this encoding to be indistinguishable from the bit-flip channel, the probability of the message being an error string of weight w should equal the value from the distribution in Eq. (1) above. This means we want to satisfy

$$qC_w = \frac{C_w}{C} = p_w. \tag{4}$$

Clearly, we must have

$$C_w \leqslant \binom{N}{w},$$

for all w in the typical range. This implies that

$$\begin{aligned} C_w p^w(1-p)^{N-w} &\leqslant \binom{N}{w} p^w(1-p)^{N-w} \qquad = C_w q \\ \Rightarrow p^w(1-p)^{N-w} &\leqslant q. \end{aligned} \tag{5}$$

To communicate the maximum amount of information steganographically, we want C to be as large as possible, which means we want q to be as small as possible. The constraint in Eq. (5) then gives us

$$q = p^{Np(1-\delta)}(1-p)^{N(1-p+p\delta)}. \tag{6}$$

So Alice can send M stegoqubits to Bob, where

$$\begin{aligned} M &= \log_2 C = \log_2 1/q \\ &= N\{-p\log_2 p - (1-p)\log_2(1-p) \\ &\quad + \delta[p\log_2 p - p\log_2(1-p)]\} \\ &= N\{h(p) - \delta p \log_2[(1-p)/p]\} \approx Nh(p), \end{aligned} \tag{7}$$

where $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the entropy of the bit-flip channel on one qubit. Therefore, with this encoding, Alice can send almost $Nh(p)$ bits. Note that in realistic scenarios, p will always be less than 0.5; otherwise the decoder can be trivially adjusted by flipping the interpretation of a 0 signal and a 1 signal.

In [7], it is shown that the diamond norm distance between the channel $(\mathcal{N}_p^{BF})^{\otimes N}$ and Alice's encoding is exponentially small in N. This justifies the claim that this protocol will not arouse suspicion from Eve. In Sec. III, we use a slightly modified definition of secrecy that allows us to prove the converse bound on this rate of stegocommunication by information theoretic techniques. That means that this encoding is essentially optimal: the maximum rate of steganographic communication for a nondegenerate code in the case of the bit-flip channel is $h(p)$.

B. Depolarizing channel

Here we will consider the scenario where the channel that Alice is emulating is the depolarizing channel. It turns out that due to the symmetric nature of the depolarizing channel, the encoding looks quite similar to that of the bit-flip channel. Recall that the depolarizing channel acting on a single qubit ρ is given by

$$\mathcal{N}_p^{\mathrm{DC}}(\rho) = (1-p)\rho + (p/3)(X\rho X + Y\rho Y + Z\rho Z).$$

Applying this channel on N qubits, the total probability of all errors with exactly n_1 X, n_2 Y, and n_3 Z errors (and $n_4 = N - n_1 - n_2 - n_3$ identity "errors") is

$$p(n_1, n_2, n_3, n_4) = \frac{N!}{n_1!n_2!n_3!n_4!}(p/3)^{n_1+n_2+n_3}(1-p)^{n_4}.$$

Notice that instead of specifying n_1, n_2, and n_3 exactly, we can instead talk about errors with weight $w = n_1 + n_2 + n_3$. It follows by simple calculation that the total probability of all errors of weight w is

$$p(w) = 3^w\binom{N}{w}(p/3)^w(1-p)^{N-w} = \binom{N}{w}p^w(1-p)^{N-w},$$

which is just a binomial distribution in w. As in the bit-flip case, we will need to say what strings of errors are typical. There are a number of ways we could specify this, but for simplicity we will consider weights w that lie between $Np(1-\delta)$ and $Np(1+\delta)$ for $\sqrt{(1-p)/pN} \ll \delta \ll 1$. The astute reader will notice that this set includes some errors that are not typical: for instance, it includes errors of weight w where all

(or most) of the errors are X's and none (or few) are Y's or Z's. If such errors are used as codewords, they might make Eve suspicious. Still, the effect of this is not too large because this set is still dominated by typical errors, and the probabilities of these strings are similar to the expected probabilities of atypical errors. For more information regarding the typicality of channel errors, we refer the reader to the Appendix. With this definition of typicality, we can follow the exact same encoding given for the bit-flip code using errors with weight w, except that the set of errors of weight w is now of size

$$\binom{N}{w}3^w,$$

and errors of weight w have probability $(p/3)^w(1-p)^{N-w}$. This then leads to the following encoding rate:

$$\begin{aligned} M &= N\{-p\log_2(p/3) - (1-p)\log_2(1-p) \\ &\quad + \delta[p\log_2(p/3) - p\log_2(1-p)]\} \\ &= N\{s(p) + \delta[p\log_2(p/3) - p\log_2(1-p)]\} \\ &\approx Ns(p), \end{aligned} \tag{8}$$

where we have defined $s(p) = -p\log_2(p/3) - (1-p)\log_2(1-p)$ to be the entropy of the depolarizing channel on one qubit.

C. General channels

1. Special case: Random unitaries

Consider a quantum channel acting on a single qubit of the form

$$\mathcal{N}(\rho) = \sum_{i=1}^{k} p_i U_i \rho U_i^\dagger, \tag{9}$$

where the operators U_i are all unitary, so $U_i U_i^\dagger = U_i^\dagger U_i = I$. The set of Kraus operators $\{\sqrt{p_i}U_i\}$ can be thought of as a set of possible single-qubit unitary errors U_i that occur with probability p_i. Note that both the bit-flip and depolarizing channels are special cases of the random unitary channel, as is any Pauli channel. The channel acts on an N-qubit encoded state ρ as $\mathcal{N}^{\otimes N}(\rho)$.

The total probability of all errors with n_1 U_1 errors, n_2 U_2 errors, and so forth is given by the multinomial distribution,

$$p(n_1,\ldots,n_k) = \frac{N!}{n_1!\cdots n_k!}p_1^{n_1}\cdots p_k^{n_k}. \tag{10}$$

Now consider weights n_j in the range from $Np_j(1-\delta)$ to $Np_j(1+\delta)$, where δ is large enough that this set includes all the typical strings. (This definition can be modified, but for simplicity we stick with it in this paper.) Randomly choose $C_{n_1,\ldots,n_k}$ strings with weights $n_1, n_2, \ldots, n_k$ in this range such that $n_1+\cdots+n_k = N$. As with the bit-flip and depolarizing channels, let these sets of strings be called $S_{n_1,\ldots,n_k}$ and let S denote the union of all these sets of strings, which are a subset of the typical strings. For all weights $n_1,\ldots,n_k$ outside the typical set, we let $C_{n_1,\ldots,n_k} = 0$. The total number of strings in the set S is C,

$$C = \sum_{n_1,\ldots,n_k} C_{n_1,\ldots,n_k}. \tag{11}$$

Defining $q \equiv 1/C$, we want to satisfy

$$C_{n_1,\ldots,n_k}q = C_{n_1,\ldots,n_k}/C = p(n_1,\ldots,n_k) \tag{12}$$

for all weights $n_1,\ldots,n_k$ in the typical set, so that Eve does not become suspicious. Also, clearly, $C_{n_1,\ldots,n_k}$ must be less than $\frac{N!}{n_1!\cdots n_k!}$. This implies that

$$\begin{aligned} &C_{n_1,\ldots,n_k}p_1^{n_1}\cdots p_k^{n_k} \leqslant \frac{N!}{n_1!\cdots n_k!}p_1^{n_1}\cdots p_k^{n_k}, \\ &C_{n_1,\ldots,n_k}p_1^{n_1}\cdots p_k^{n_k} \leqslant C_{n_1,\ldots,n_k}q, \\ &p_1^{n_1}\cdots p_k^{n_k} \leqslant q. \end{aligned} \tag{13}$$

Notice that this time we cannot simply plug in the lower bounds of the sums for n_j, as we did for the depolarizing and bit-flip channels, because we have the additional constraint that $n_1+\cdots+n_k = N$. However, the same general argument applies. Inside the set of typical weights, there is a string $\tilde{n}_1,\ldots,\tilde{n}_k$ with $|\tilde{n}_j/N - p_j| \leqslant \delta p_j$ for all j, which maximizes the probability,

$$p_{\max} \equiv p_1^{\tilde{n}_1}p_2^{\tilde{n}_2}\cdots p_k^{\tilde{n}_k}. \tag{14}$$

We can choose $q = p_{\max}$ and use this to put a bound on the number of stegoqubits M that Alice can send to Bob,

$$\begin{aligned} M &= \log_2 C = -\log_2(q) = -\log_2 p_{\max} \\ &= -\tilde{n}_1\log_2(p_1) - \cdots - \tilde{n}_k\log_2(p_k) \\ &= N\left[-\frac{\tilde{n}_1}{N}\log_2(p_1) - \cdots - \frac{\tilde{n}_k}{N}\log_2(p_k)\right] \\ &\geqslant N(1-\delta)\left[-\sum_{i=1}^{k} p_i\log_2(p_i)\right] \\ &= N(1-\delta)H(p_1,\ldots,p_k). \end{aligned} \tag{15}$$

So, in the limit of large N, we should approach a rate $H(p_1,\ldots,p_k)$ with this encoding.

2. Encoding general channels across multiple code blocks

This argument does not necessarily apply directly to a general quantum channel since the probabilities of the different outcomes can be state dependent. However, we should be able to do a similar type of encoding for a general quantum channel $\mathcal{N}$ by encoding across multiple code blocks. Consider a general quantum channel acting on a single qubit as

$$\mathcal{N}(\rho) = \sum_{i=1}^{k} A_i \rho A_i^\dagger. \tag{16}$$

The channel acts on an N-qubit encoded state ρ as $\mathcal{N}^{\otimes N}(\rho)$, where we will let N become large. For most states ρ, we can well approximate this N-qubit channel by a sum over the *typical* errors [16,17],

$$\mathcal{N}^{\otimes N}(\rho) \approx \sum_{\underline{i}\in\mathcal{T}} E_{\underline{i}}\rho E_{\underline{i}}^\dagger, \tag{17}$$

where ρ is now the N-qubit codeword, the index is $\underline{i} = i_1 i_2 \cdots i_N$, the typical error $E_{\underline{i}}$ is

$$E_{\underline{i}} = A_{i_1}\otimes A_{i_2}\otimes\cdots\otimes A_{i_N}, \tag{18}$$

and $\mathcal{T}$ is the set of typical sequences $\underline{i}$ [18].

We assume that the QECC that Alice uses is one that can correct the typical errors of the channel. (Indeed, using a code that was not strong enough to correct the typical errors might well arouse Eve's suspicions.) We will also assume, for simplicity of this analysis, that the QECC is *strongly nondegenerate*.

Definition. Strongly nondegenerate code. Suppose a quantum channel $\mathcal{N}$ acts on a single qubit as follows:

$$\mathcal{N}(\rho) = \sum_l A_l \rho A_l^\dagger. \tag{19}$$

Then we say a quantum error-correcting code is strongly nondegenerate if there exists a correctable set of errors of the form

$$E_{\underline{i}} = A_{l_1(i)} \otimes A_{l_2(i)} \otimes \cdots \otimes A_{l_n(i)}, \tag{20}$$

such that $\mathbb{P} E_i^\dagger E_j \mathbb{P} = \alpha_{ij} \mathbb{P}$, where $\mathbb{P}$ is the codespace projector and α_{ij} is diagonal.

This means that on a valid codeword in the QECC, the typical errors $E_{\underline{i}}$ all have distinct error syndromes and act as unitaries that move the state to a distinct, orthogonal subspace labeled by $\underline{i}$. This means that error $E_{\underline{i}}$ occurs with a fixed probability $p_{\underline{i}}$ for all valid codewords of the QECC.

We can then essentially repeat the argument that leads to Eq. (15), but now using the probabilities $p_{\underline{i}}$. Note that we now need to take two limits: the limit of many blocks and the limit where the individual blocks are large. For this argument to apply, we need to first go to the limit of many blocks and then to the limit of large block size. In those limits, we can approach a rate

$$-\frac{1}{N} \sum_{\underline{i}} p_{\underline{i}} \log_2 p_{\underline{i}} \equiv \bar{H}, \tag{21}$$

where $\bar{H}$ is an effective entropy per qubit from the channel. It is likely that this strong definition of code nondegeneracy is not needed and Eq. (21) will still hold under a weaker definition; this will be the topic of future work.

D. Secret key consumption

For the above encodings, how much of the secret key must be consumed? In general, we can assume that all the details of the encoding, etc. have been decided between Alice and Bob ahead of time. So in the protocol as described above, the only place where the secret key is consumed is to pick the subsets of errors used in the encoding.

Let us consider the bit-flip channel as a simple example. The possible messages are mapped onto a set of C error syndromes, representing errors of weights $(1-\delta)Np \leqslant w \leqslant (1+\delta)Np$. For each error weight w in that range, a subset of C_w errors is chosen to represent possible messages. Alice and Bob can agree before the protocol begins to divide the set of errors of weight w into n_w nonoverlapping subsets of C_w errors each, where

$$n_w = \binom{N}{w} \Big/ C_w = \left(\frac{1-p}{p}\right)^{w-Np(1-\delta)}. \tag{22}$$

(Since this is unlikely to be an exact integer, one must generally round down, which means that a small fraction of possible errors will be omitted. This will slightly reduce the match between the steganographic encoding and the noisy channel being simulated, but for large N and $p \ll 1$, the difference will be small.)

For each transmitted block, Alice and Bob must randomly choose one of these n_w subsets for each weight w in the typical range. Choosing a subset requires $\log_2 n_w$ random bits, which are drawn from their shared key. However, since any given message is encoded as an error of some specific weight w, Alice and Bob can reuse the same secret key bits to choose the subset for each error weight w. So the number of key bits consumed to transmit one block is equal to the maximum value of $\log_2 n_w$ for $(1-\delta)Np \leqslant w \leqslant (1+\delta)Np$, which is

$$\begin{aligned} K &= \max_{Np(1-\delta)\leqslant w\leqslant Np(1+\delta)} \log_2 n_w \\ &= \max_{Np(1-\delta)\leqslant w\leqslant Np(1+\delta)} \log_2 \left(\frac{1-p}{p}\right)^{w-Np(1-\delta)} \\ &= (2Np\delta)\log_2\left(\frac{1-p}{p}\right). \end{aligned} \tag{23}$$

How does this scale with N? Since this is a binomial distribution, δ will take the form

$$\delta = D\sqrt{\frac{1}{N}\left(\frac{1-p}{p}\right)}, \tag{24}$$

where D is a fixed constant determining what fraction of all errors are included in the typical set. The key consumption therefore is

$$K = 2D\sqrt{N\left(\frac{1-p}{p}\right)}\log_2\left(\frac{1-p}{p}\right). \tag{25}$$

The key consumption scales sublinearly with N, and asymptotically the key consumption rate goes to zero. While the details will vary, we expect this kind of sublinear scaling of K with N to be generic. Note that most of the randomness in Alice's preparation of the state can be generated locally. Only the randomness used in choosing the particular encoding needs to be shared with Bob, and so it is important that Alice and Bob share the key beforehand so that Bob knows which set of errors is being used to encoded messages.

A few words more on secret key consumption are in order. In [7], Shaw and Brun make a distinction between the *secrecy* and the *security* of a steganographic protocol. A steganographic protocol is *secret* if an eavesdropper without the secret key cannot distinguish between an encoded message being sent and the noisy channel being applied. It is *secure* if the eavesdropper cannot learn anything about the message, even if she knows that a message is begin sent.

Using a sublinear amount K of the shared secret key is sufficient to make the steganographic protocol secret, by this definition. However, it is *not* secure, in general. Since the number of qubits M transmitted is typically larger than the number of secret key bits K consumed, we would generically expect an eavesdropper to be able to learn of the order of $M-K$ bits of information about the message if she became aware of its existence.

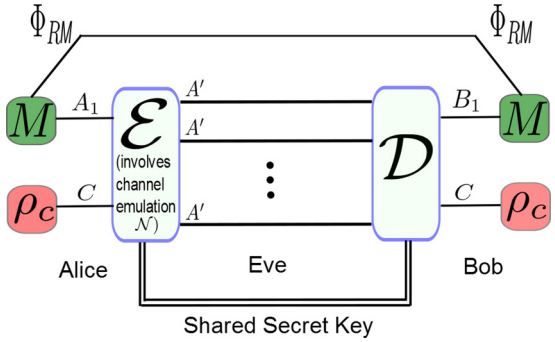

FIG. 1. The information-processing task that we consider for Alice sending M stegoqubits to Bob over a quantum channel (which is identity for the noiseless case). Alice encodes her message M and an innocent covertext ρ_c into a suitable quantum error-correcting code which has had typical errors applied to it, where the encoding depends on the secret key k. She sends this to Bob, who then decodes the message and covertext using his copy of the shared secret key k. Alice's message is entangled with a reference system R. The ability to transmit entanglement implies the ability to do general quantum communication.

This can be prevented by first encrypting the message before doing the steganographic encoding. Encryption requires M bits of the secret key in the case of a classical message (using a one-time pad) or $2M$ bits of the secret key in the case of a quantum message (by twirling). In this case, the protocol is both secret *and* secure. However, there is a cost: the secret key is now consumed asymptotically at a linear rate.

III. SECRECY, RELIABILITY, AND BOUNDS

A. The information-processing task

Here we consider the steganographic scenario as outlined above where Alice is using fake noise to hide her message from Eve, but the actual physical channel she is sending her information over is noiseless. We will consider the task known as *entanglement transmission*. This notion of quantum communication encompasses other quantum information-processing tasks such as mixed-state transmission, pure-state transmission, and entanglement generation. We closely follow the discussion of quantum communication in [18].

The information-processing task that we are considering is visualized in Fig. 1. Alice has a secret message of $M = \log_2 |A_1|$ qubits, which is maximally entangled with a reference system R. She also prepares an innocent covertext ρ_c, which will be encoded into the N-qubit quantum error-correcting code. Let us first define her encoded state, dependent on the secret key element k,

$$\omega_{k,A'^n R} \equiv \mathcal{E}_{k,A_1 C \to A'^n}(\rho_c \otimes \Phi_{A_1 R}). \tag{26}$$

This dependence of the encoding on the secret key corresponds to choosing among the different sets of error strings S in the protocols from the previous section. To someone (such as Eve) who does not know the secret key k, the state is effectively

$$\omega_{A'^n R} \equiv \sum_k p_k \omega_{k,A'^n R}, \tag{27}$$

where $\omega_{A'^n R}$ is the state averaged over all possible values of the secret key k with probabilities p_k. (We can choose this probability to be uniform for simplicity, $p_k = p$ for all k, if we so desire.)

What is a good way to guarantee secrecy from Eve? We propose the following *secrecy* condition:

$$\tfrac{1}{2} \| \operatorname{Tr}_R(\omega_{A'^n R}) - \mathcal{N}^{\otimes N}(V \rho_c V^\dagger) \|_1 \leqslant \delta, \tag{28}$$

where $\mathcal{N}$ is whatever channel Alice is emulating, V is an isometry representing the encoding of the covertext into a suitably chosen codeword (one which can correct typical errors induced by the channel $\mathcal{N}$), and $\delta > 0$ is some small parameter. What this condition says is that if Eve observes the quantum state, it will be effectively indistinguishable from an encoded covertext being sent through the noisy quantum channel $\mathcal{N}$.

We introduce another requirement which corresponds to a notion of *recoverability*. Once Bob receives the state, he applies his decoder $\mathcal{D}_{k,A'^n \to B_1 C}$ to obtain the original $\rho_c \otimes \Phi_{B_1 R}$. We can relax this by only requiring that the input states and output states are ϵ close, that is,

$$\tfrac{1}{2} \| \mathcal{D}_{k,A'^n \to B_1 C}(\omega_{k,A'^n R}) - \rho_c \otimes \Phi_{B_1 R} \|_1 \leqslant \epsilon, \forall k, \tag{29}$$

where $\epsilon > 0$ is a small parameter.

B. Upper bound on steganographic rate

With these two assumptions of secrecy and recoverability, we can now put a bound on the number of qubits M that can be sent reliably and steganographically from Alice to Bob. Defining $\sigma_E \equiv \mathcal{N}^{\otimes N}(V \rho_c V^\dagger)$ and applying the Fannes-Audeneart inequality to the secrecy condition, we have

$$H(\operatorname{Tr}_R(\omega_{A'^n R})) \leqslant H(\sigma_E) + \delta N + h_2(\delta), \tag{30}$$

where h_2 is the binary entropy function. Furthermore, from the recoverability condition, we have

$$\begin{aligned} M &= \log_2 |A_1| = I(R\rangle B_1)_\Phi \\ &\leqslant I(R\rangle B_1)_{\mathcal{D}_k(\omega)} + \epsilon N + (1+\epsilon) h_2(\epsilon/[1+\epsilon]) \\ &\leqslant I(R\rangle A'^n)_{\omega_k} + f(N,\epsilon) \\ &\leqslant H(\operatorname{Tr}_R(\omega_{k,A'^n R})) + f(N,\epsilon), \end{aligned} \tag{31}$$

where $I(R\rangle B_1)_\Phi$ denotes the quantum coherent information between the registers R and B_1 in the quantum state Φ_{RB_1}. The first equality follows from the fact that the coherent information of a maximally entangled state is just the logarithm of the dimension of one of the subsystems. The first inequality follows from the Alicki-Fannes-Audeneart inequality applied to (29). The second inequality is the data-processing inequality. The last inequality follows from the definition of the coherent information. We point the reader to the Appendix for definitions of these quantities and inequalities.

The concavity of entropy implies that

$$\sum_k p_k H(\omega_{k,A'^n}) \leqslant H\left(\sum_k p_k \omega_{k,A'^n} \right) = H(\omega_{A'^n}). \tag{32}$$

The encodings $\mathcal{E}_{k,A_1 C \to A'^n}$ are isometries, which means that $H(\omega_{k,A'^n})$ has the same value for every k. We can therefore

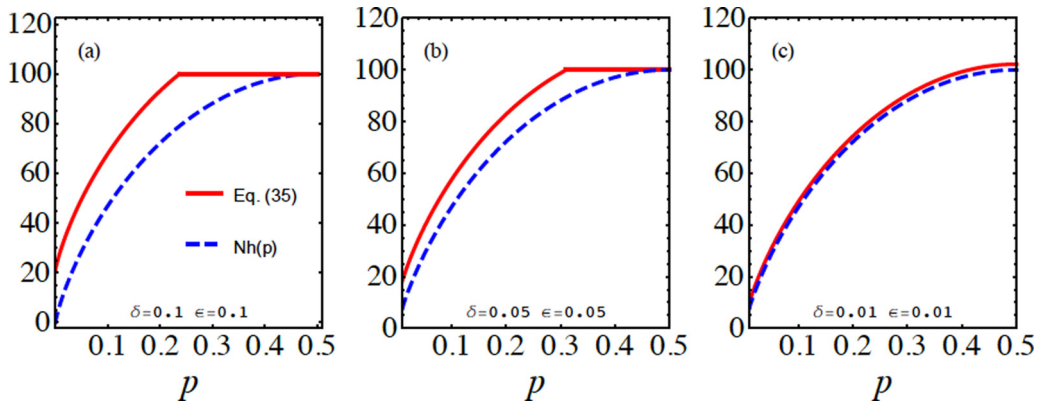

FIG. 2. Here we plot the number of classical or quantum bits M sent vs the error rate of the bit-flip channel (both dimensionless). The solid red curve is the upper bound on the number of classical or quantum bits M that Alice can send to Bob steganographically over an $N = 100$ qubit block, given by Eq. (37). The dashed blue curve is the limit of the achievable rate for the bit-flip channel given by $Nh(p)$. These two curves are plotted for three different values of the secrecy and reliability parameters for steganography. The secrecy parameter δ which is set by Eq. (28) determines how likely Eve is to become suspicious, and the reliability parameter ϵ which is set by Eq. (29) describes how reliable Bob's decodings of the messages are. As can be seen from these plots, the more secret and reliable the communication protocol, the more the upper bound and achievable rate coincide.

sum over the probabilities p_k on the left-hand side of (32) to get

$$H(\mathrm{Tr}_R(\omega_{k,A^nR})) \leqslant H(\mathrm{Tr}_R(\omega_{A^nR})). \tag{33}$$

Now putting (30) and (31) together, we arrive at our main result, which states that Alice can secretly and reliably send M stegoqubits to Bob, where M is bounded above by

$$\begin{aligned} M &\leqslant H(\mathrm{Tr}_R(\omega_{RA^n})) + f(N,\epsilon) \\ &\leqslant H(\sigma_E) + g(N,\delta) + f(N,\epsilon), \end{aligned} \tag{34}$$

where $g(N,\delta) \equiv \delta N + h_2(\delta)$, and $f(N,\epsilon) \equiv \epsilon N + (1+\epsilon)h_2(\epsilon/[1+\epsilon])$. Thus, if we can compute a maximum for $H(\mathcal{N}^{\otimes N}(\rho))$ when ρ is pure (because V is an isometric encoding and ρ_c is pure), we have a tight upper bound on the number of qubits M that can be sent steganographically over a noiseless quantum channel. (Of course, if the actual quantum channel is noisy, then this bound will in general be changed. This is the topic of future work.)

C. Upper bounds for specific channels

We will now apply our result (34) to the channels discussed in the previous section, where we make the implicit assumption that Alice is using a nondegenerate code. Though our result (34) is true in general, for a degenerate code the number of distinct error syndromes is smaller (depending on the code) and the bounds discussed here and achievable rates discussed in the previous section would be adjusted.

1. The bit-flip channel

For the bit-flip channel, i.e., $\mathcal{N}_{\mathrm{BF}}(\rho) = (1-p)\rho + pX\rho X$, the maximum of $H(\mathcal{N}^{\otimes N}(\rho))$ over all N-qubit pure states ρ is $Nh(p)$, where $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the entropy of a single qubit sent through a bit-flip channel. To prove this, consider some pure state $\rho = |\psi\rangle\langle\psi|$. Then,

$$\mathcal{N}_{\mathrm{BF}}^{\otimes N}(|\psi\rangle\langle\psi|) = \sum_s p(s) X^s |\psi\rangle\langle\psi| X^s, \tag{35}$$

where we are summing over all binary strings s of length N; X^s is the operator acting on N qubits with an X acting at every location where s has a 1 and an I where s has a 0. The probability $p(s)$ is given by

$$p(s) = p^{w(s)}(1-p)^{[N-w(s)]}, \tag{36}$$

where $w(s)$ is the weight of string s. The Shannon entropy of this distribution is $Nh(p)$ since it is a binomial distribution. The von Neumann entropy is the minimum Shannon entropy over all possible ensemble decompositions of the given state, and it is not hard to check that it is achieved when $|\psi\rangle$ is a Z eigenstate. Therefore, for the bit-flip channel, Eq. (34) becomes

$$M \leqslant Nh(p) + (\delta+\epsilon)N + (1+\epsilon)h_2(\epsilon/[1+\epsilon]) + h_2(\delta), \tag{37}$$

and thus the encoding described in the previous section for steganography with a simulated bit-flip channel is essentially optimal. As can be seen from the plot of Eq. (37) in Fig. 2, it converges to the achievable limit $Nh(p)$ as the secrecy and reliability parameters approach 0.

2. More general channels

Unfortunately, for a more general quantum channel $\mathcal{N}$, we may not know, in general, what N-qubit pure-state ρ maximizes $H(\mathcal{N}^{\otimes N}(\rho))$. However, we can still bound this quantity. First, consider a general quantum channel $\mathcal{N}$ that acts on an N qubit pure state as follows:

$$\mathcal{N}^{\otimes N}(\rho) \approx \sum_j E_j \rho E_j^\dagger, \tag{38}$$

where $\{E_j\}$ is the set of typical errors associated with N applications of the channel $\mathcal{N}$. Recall that we are choosing our isometric encoding to correct for typical errors of whatever channel $\mathcal{N}$ it is that we are emulating. Though the set of correctable errors $\{E_j\}$ need not act like unitaries on the codespace, we can always find a set of correctable errors $\{\widetilde{E}_j\}_j$ that do [19]. To see this, first consider the Knill-Laflamme

condition,

$$\mathbb{P}E_i^\dagger E_j \mathbb{P} = \alpha_{ij}\mathbb{P}, \tag{39}$$

where $\mathbb{P}$ is the codespace projector and α is a Hermitian matrix. Thus, we can write $\widetilde{\alpha} = U^\dagger \alpha U$, where U is a unitary matrix and $\widetilde{\alpha}$ is diagonal.

$$\widetilde{E}_k = \sum_j M_{jk} E_k, \tag{40}$$

where the unitary M is chosen in such a way as to diagonalize α. That is,

$$\begin{aligned}\mathbb{P}\widetilde{E}_k^\dagger \widetilde{E}_l \mathbb{P} &= \sum_{ij} M_{ik}^* M_{jl} \mathbb{P}E_i^\dagger E_j \mathbb{P} = \left(\sum_{ij} M_{ik}^* \alpha_{ij} M_{jl}\right)\mathbb{P} \\ &= \widetilde{\alpha}_{kl}\mathbb{P} = \delta_{kl}\widetilde{\alpha}_{kk}\mathbb{P}.\end{aligned} \tag{41}$$

Note that these errors $\{\widetilde{E}_j\}$ act unitarily on the codespace. So long as the Knill-Laflamme condition is satisfied, we can always diagonalize α in this way. Now going back to our expression for the channel action,

$$\sum_j E_j \rho E_j^\dagger = \sum_{k,l,j} M_{kj} M_{lj}^* \widetilde{E}_k \rho \widetilde{E}_l^\dagger = \sum_k \widetilde{E}_k \rho \widetilde{E}_k^\dagger. \tag{42}$$

Because we have assumed that the typical errors are all correctable and the code is nondegenerate, the states $\widetilde{E}_k \rho \widetilde{E}_k^\dagger$ are all orthogonal to each other, and $\mathrm{Tr}\{\widetilde{E}_k \rho \widetilde{E}_k^\dagger\} = \alpha_{kk}$. The von Neumann entropy is the Shannon entropy minimized over all possible decompositions, so the entropy of this state is clearly

$$H(\sigma_E) = H(\mathcal{N}^{\otimes N}(V\rho_c V^\dagger)) \leqslant -\sum_k \alpha_{kk}\log_2(\alpha_{kk}). \tag{43}$$

By (34), we have shown that the amount of steganographic communication allowed for a quantum channel $\mathcal{N}$ emulation is upper bounded by this quantity. Applying this to the general channel discussed in Sec. II C above, we see that this quantity is equal to $N\bar{H}$, where $\bar{H}$ is the effective entropy per qubit defined in Eq. (21). So this encoding approaches the maximum possible rate for the general channel, just as for the bit-flip channel.

IV. CONCLUSIONS AND FUTURE WORK

Quantum steganography is the study of secret quantum communication between two parties, Alice and Bob. We have shown that Alice and Bob are able to communicate with each other secretly at a nonzero rate using a shared secret key, without arousing suspicion from a potential eavesdropper Eve. In this paper, we gave explicit bounds on the number of stego-qubits that Alice can send to Bob when Alice is simulating a general quantum channel $\mathcal{N}$ with her stego encoded message, as well as explicit encodings to achieve these bounds, for the case when the actual physical channel is noiseless.

The obvious next question is what if the channel shared between Alice and Bob (as is generally the case) is noisy? There is reason to believe that as long as Eve has some ignorance about the actual physical channel, Alice will still be able to communicate steganographically to Bob.

For instance, suppose the actual physical channel is a depolarizing channel $\mathcal{N}_p$, where p is the depolarizing parameter and the channel that Eve expects is $\mathcal{N}_{p+\epsilon-4p\epsilon/3}$ for some small suitably chosen $\epsilon > 0$. Then Alice can emulate a depolarizing channel $\mathcal{N}_\epsilon$ in such a way such that if Eve observes the state Alice is sending to Bob, it will look like an innocent encoded covertext passing through N applications of a channel $\mathcal{N}_p \circ \mathcal{N}_\epsilon$ (where N is the length of the codeword Alice is using). There should be elements of the encoding given in this paper that will generalize to the noisy case for general channels $\mathcal{N}$. This will certainly be an area of fruitful future study.

ACKNOWLEDGMENTS

The authors thank Mark Wilde and David Ding for helpful discussions. This research was supported in part by NSF Grants No. CCF-1421078 and No. QIS-1719778, and by an IBM Einstein Fellowship at the Institute for Advanced Study.

APPENDIX: TYPICAL CHANNEL ERRORS, ENTROPY, AND INEQUALITIES

Here we give the more explicit arguments related to our heuristic arguments in Sec. II about typical errors and effective entropies. First we give the definition of typical errors associated with a channel $\mathcal{N}^{\otimes N}$ as originally outlined in [16,17]. Consider a quantum channel $\mathcal{N}$ with Kraus operators $E_1, \ldots, E_N$. Without loss of generality, we can assume they are diagonal, i.e., $\mathrm{Tr}\, E_i E_j^\dagger = 0$ for $i \neq j$. The probability of each Kraus operator is given by $p_i = \frac{1}{M}\mathrm{Tr}\, E_i E_i^\dagger$, where M is the dimension of the Hilbert space on which these operators act.

Now the operator $\mathcal{N}^{\otimes N}$ can be represented by N^m Kraus operators,

$$E_{j_1} \otimes E_{j_2} \otimes \cdots \otimes E_{j_n} \equiv E_J, \tag{A1}$$

where $j_i = 1, \ldots, N$. It is straightforward to verify that the probability associated with each of these operators E_J is given by

$$p_J = \frac{1}{M^n}\mathrm{Tr}\, E_J^\dagger E_J = p_{j_1}\ldots p_{j_n}. \tag{A2}$$

Hence, the Kraus operators E_J of $\mathcal{N}^{\otimes N}$ are sequences of length n in which the symbols E_i of an alphabet $E_1, \ldots, E_N$ appear according to the probability distribution $\{p_i\}$. Hence we are in the domain of classical random sequences. Thus we can take only the operators E_J that are ϵ-typical in the usual sense with respect to this probability distribution, and write

$$\mathcal{N}^{\otimes N}(\rho) \approx \sum_{J\ \text{typical}} E_J \rho E_J^\dagger. \tag{A3}$$

This strongly reduces the number of Kraus operators of $\mathcal{N}^{\otimes N}$ from N^m to roughly $2^{NH(\{p_i\})}$.

We now define the various inequalities which are used in Sec. III to prove our converse theorem.

Definition 1. Fannes-Audeneart inequality. Let $\rho, \sigma \in \mathcal{D}(\mathcal{H})$ be density operators and suppose that $\frac{1}{2}||\rho - \sigma||_1 \leqslant \epsilon \in [0, 1]$. Then, the following inequality holds:

$$|H(\rho) - H(\sigma)| \leqslant \epsilon \log_2 \dim(\mathcal{H}) + h_2(\epsilon).$$

Definition 2. Alicki-Fannes-Audeneart inequality. Let $\rho_{AB}, \sigma_{AB} \in \mathcal{D}(\mathcal{H}_A \otimes \mathcal{H}_B)$. Suppose that

$$\tfrac{1}{2}||\rho_{AB} - \sigma_{AB}||_1 \leqslant \epsilon,$$

for $\epsilon \in [0, 1]$. Then,

$$|H(A|B)_\rho - H(A|B)_\sigma| \leqslant 2\epsilon \log_2 \dim(\mathcal{H}_A) + g_2(\epsilon),$$

where $g_2(\epsilon) \equiv (\epsilon + 1)\log_2(\epsilon + 1) - \epsilon \log_2(\epsilon)$.

Before describing the data processing inequality used in the main text, we first define the coherent information:

Definition 3. Quantum coherent information. The coherent information $I(A\rangle B)_\rho$ of a bipartite state $\rho_{AB} \in \mathcal{D}(\mathcal{H}_A \otimes \mathcal{H}_B)$ is given by

$$I(A\rangle B)_\rho \equiv H(B)_\rho - H(AB)_\rho,$$

where $H(\rho)$ is the von Neumann entropy.

The quantum coherent information satisfies the following inequality:

Definition 4. Data processing for coherent information. Let $\rho_{AB} \in \mathcal{D}(\mathcal{H}_A \otimes \mathcal{H}_B)$ and let $\mathcal{N} : \mathcal{L}(\mathcal{H}_B) \to \mathcal{L}(\mathcal{H}_{B'})$ be a quantum channel. Set $\sigma_{AB'} \equiv \mathcal{N}_{B\to B'}(\rho_{AB})$. Then the following quantum data-processing inequality holds:

$$I(A\rangle B)_\rho \geqslant I(A\rangle B')_\sigma.$$

[1] Herodotus, *The Histories* (Penguin Books, London, United Kingdom, 1996).

[2] *Velvalee Dickinson, the "Doll Woman"*, https://www.fbi.gov/history/famous-cases/velvalee-dickinson-the-doll-woman (unpublished).

[3] M. Dušek, N. Lütkenhaus, and M. Hendrych, Prog. Opt. **49**, 381 (2006).

[4] S. Natori, in *Quantum Computation and Information* (Springer, New York, 2006), pp. 235–240.

[5] I. Banerjee, S. Bhattacharyya, and G. Sanyal, Intl. J. Comput. Network Inf. Secur. **4**, 65 (2012).

[6] J. Gea-Banacloche, J. Math. Phys. **43**, 4531 (2002).

[7] B. A. Shaw and T. A. Brun, Phys. Rev. A **83**, 022310 (2011).

[8] B. A. Shaw and T. A. Brun, arXiv:1007.0793.

[9] B. A. Bash, A. H. Gheorghe, M. Patel, J. L. Habif, D. Goeckel, D. Towsley, and S. Guha, Nat. Commun. **6**, 8626 (2015).

[10] A. Sheikholeslami, B. A. Bash, D. Towsley, D. Goeckel, and S. Guha, in *2016 IEEE International Symposium on Information Theory (ISIT)*, edited by D. Abbott (IEEE, Barcelona, 2016), pp. 2064–2068.

[11] L. Wang, in *2016 IEEE Information Theory Workshop (ITW)*, edited by D. Abbott (IEEE, Cambridge, UK, 2016) pp. 364–368.

[12] K. Bradler, T. Kalajdzievski, G. Siopsis, and C. Weedbrook, arXiv:1607.05916.

[13] J. M. Arrazola and V. Scarani, Phys. Rev. Lett. **117**, 250503 (2016).

[14] S. P. Desrosiers, Quantum Inf. Proc. **8**, 331 (2009).

[15] Y. Dodis and A. Smith, in *Theory of Cryptography Conference*, edited by Y. Lindell (Springer, New York, 2005), pp. 556–577.

[16] R. Klesse, Phys. Rev. A **75**, 062315 (2007).

[17] R. Klesse, Open Syst. Inf. Dyn. **15**, 21 (2008).

[18] M. M. Wilde, *Quantum Information Theory* (Cambridge University Press, Cambridge, 2013).

[19] M. A. Nielsen and I. Chuang, Am. J. Phys. **70**, 558 (2020).