Quantum steganography with noisy quantum channels

Bilal A. Shaw[1,2,*] and Todd A. Brun[1,2,3,4,†]

[1]*Department of Computer Science, University of Southern California, Los Angeles, California 90089, USA*
[2]*Center for Quantum Information Science and Technology, University of Southern California, Los Angeles, California 90089, USA*
[3]*Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089, USA*
[4]*Department of Physics and Astronomy, University of Southern California, Los Angeles, California 90089, USA*
(Received 10 November 2010; published 14 February 2011)

Steganography is the technique of hiding secret information by embedding it in a seemingly "innocent" message. We present protocols for hiding quantum information by disguising it as noise in a codeword of a quantum error-correcting code. The sender (Alice) swaps quantum information into the codeword and applies a random choice of unitary operation, drawing on a secret random key she shares with the receiver (Bob). Using the key, Bob can retrieve the information, but an eavesdropper (Eve) with the power to monitor the channel, but without the secret key, cannot distinguish the message from channel noise. We consider two types of protocols: one in which the hidden quantum information is stored locally in the codeword, and another in which it is embedded in the space of error syndromes. We analyze how difficult it is for Eve to detect the presence of secret messages, and estimate rates of steganographic communication and secret key consumption for specific protocols and examples of error channels. We consider both the case where there is no actual noise in the channel (so that all errors in the codeword result from the deliberate actions of Alice), and the case where the channel is noisy and not controlled by Alice and Bob.

DOI: 10.1103/PhysRevA.83.022310 PACS number(s): 03.67.Dd, 03.67.Hk, 03.67.Pp

I. INTRODUCTION

Steganography is the science of hiding a message within a larger innocent-looking plain-text message and communicating the resulting data over a communications channel or by a courier so that the steganographic message is readable only by the intended receiver. The word comes from the Greek words *steganos* which means "covered," and *graphia* which means "writing." The art of information hiding dates back to 440 B.C. to the Greeks [1]. The term steganography was first used in 1499 by Johannes Trithemius in his *Steganographia*, which was one of the first treatises on the use of cryptographic and steganographic techniques [2].

The modern study of steganography was initiated by Simmons, and the paradigm can be stated as follows [3]: Alice and Bob are imprisoned in two different cells that are far apart. They would like to devise an escape plan, but the only way they can communicate with each other is through a courier who is under the command of the warden (Eve, the adversary) of the penitentiary. The courier leaks all information to the warden. If the warden suspects that either Alice or Bob are conspiring to escape from the penitentiary, she will cut off all communication between them, and move both of them to a maximum security cell. Prior to their incarceration Alice and Bob had access to a shared secret key—assumed to be a sufficiently long string of random bits—which they later exploit to send secret messages hidden in a cover text. Can Alice and Bob devise an escape plan without arousing the suspicion of the warden?

Julio Gea-Banacloche [4] introduced the idea of hiding secret messages in the form of error syndromes by deliberately applying correctable errors to a quantum state encoded in the three-bit repetition quantum error-correcting code (QECC). In his paper, however, he did not address the issue of an innocent-looking message—in the protocol he proposed, the messages would not resemble a plausible quantum channel. The latter is one of the major contributions of our work. Curty *et al.* propose three different quantum steganographic protocols [5]. However, none of these protocols address the issue of communicating an innocent message over a noisy classical channel or a general quantum channel or give key-consumption rates. Natori provides a simple treatment of quantum steganography that is a modification of super-dense coding [6]. Martin also introduced a notion of quantum steganographic communication [7]. His protocol is a variation of Bennett and Brassard's quantum-key distribution protocol (QKD), in which he hides a steganographic channel in the QKD protocol.

There are two major goals in quantum steganography. **Communication**: Alice wants to send classical or quantum information to Bob over a quantum channel. While most of the protocols above are for sending classical information, we will show that it is also possible to send quantum information. **Secrecy**: a monitoring Eve should be unable to detect the presence of the secret message. Ideally, a protocol should maximize the rate of communication as much as possible, consistent with the secrecy requirement. A third requirement may or may not be imposed: **Security**. That is, in some cases we may require that Eve be unable to read the steganographic message even if she knows it is present.

In this paper, we present a group of protocols that achieve the above goals. These protocols have the following structure. An "innocent" quantum message $|\phi_c\rangle$ is encoded in a quantum error-correcting code (QECC) by Alice. This $|\phi_c\rangle$ is the covertext. Alice then performs a second operation on the encoded covertext, which embeds the steganographic message in the codeword. This steganographic message is

*bilalsha@usc.edu
†tbrun@usc.edu

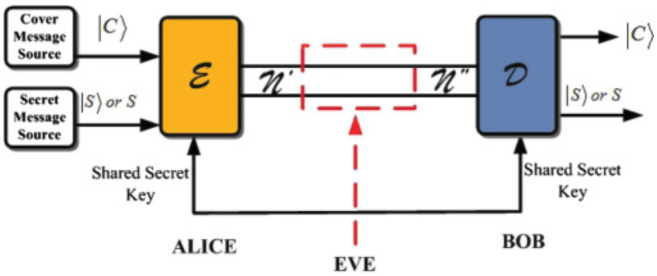

FIG. 1. (Color online) There are three different inputs to the steganographic encoder $\mathcal{E}$: a cover-message $|C\rangle$; the secret message that we would like to hide, which can be quantum $|S\rangle$ or classical S; a shared secret key which may be quantum (ebit) $|\mathcal{K}\rangle$ or classical $\mathcal{K}$. Eve can monitor some part of the noisy quantum channel $\mathcal{N}$ shown in the red box. Bob can decode the steganographic message using the decoder $\mathcal{D}$ and the shared secret key $|\mathcal{K}\rangle$ or $\mathcal{K}$ and recover $|C\rangle$ and $|S\rangle$ or S with very high probability.

another state $|\phi_s\rangle$ and is called the stego text. (We call one bit or qubit of the stego text a stego bit or stego qubit, respectively.) The modified codeword is sent over a quantum channel to Bob, who can (at least with high probability) decode it and extract the stego text $|\phi_s\rangle$. The encoding is done in such a way that, if an eavesdropper Eve intercepts the codeword, it will look exactly like the encoded state $|\phi_c\rangle$ after it has passed through a noisy channel. In other words, Eve cannot distinguish the encoded steganographic message from noise in the channel. We depict the general quantum steganographic protocol in Fig. 1.

To prove the efficacy of these protocols, we need to make a number of assumptions about the knowledge that Alice, Bob, and Eve have, and about the resources on which they can draw. These assumptions are:

(1) Alice and Bob know (with reasonable accuracy) the physical channel, which may or may not have intrinsic noise (we consider both cases). This is not unreasonable. If we imagine quantum channels are constructed from, for example, optical fiber cables, then Alice and Bob can have acquired the parameters of the cable from the manufacturer's website.

(2) Eve has beliefs about the physical channel, which may or may not be accurate. But we assume that Alice and Bob have some knowledge of Eve's expectations. This is most plausible when both Eve and Alice and Bob all draw their knowledge from the same source, but it could hold in other cases as well (e.g., the channel is actually noiseless, but Alice and Bob have systematically fooled Eve into thinking it is noisy).

(3) Alice and Bob share a secret key or shared entanglement. A secret key is a long binary string drawn from a random distribution. Shared entanglement can be used to generate such a key, but can also be used as a quantum resource for teleportation and other quantum information protocols.

(4) Eve can make measurements of any message that passes on the channel, although she will not necessarily always do so. If Eve intercepts a message, she can demand from Alice and Bob information about the covertext $|\phi_c\rangle$, the QECC used, etc., and make measurements to verify their information.

It is important to appreciate the difference in paradigms between steganography and standard cryptography. In standard cryptography, the eavesdropper is assumed to operate secretly and (perhaps) illegitimately. In steganography, the eavesdropper can operate openly, and is often in a position of authority. Eve could prevent secret communication by the simple expedient of banning all communication. But generally, she wishes to allow certain kinds of approved communication, while banning others. Cryptography is a defense against spies; steganography, against censors and secret police.

The protocols that we present in this paper succeed if Alice and Bob can communicate a nonzero amount of information, while satisfying the secrecy requirement. This demands that if Eve intercepts the message she is unable to ascertain if it contains secret information with high probability. The message should appear just like a codeword for $|\phi_c\rangle$ that has undergone some plausible set of errors in the channel. If we choose, we can further demand that even after knowing that the message contains secret information Eve will be unable to read it. This can be achieved by adding additional encryption. Alice and Bob want to maximize the rate at which they send steganographic information to each other while minimizing key usage, subject to the secrecy condition. We present protocols to achieve these goals in this paper, although we do not claim that these protocols are necessarily optimal.

We begin by giving a simple steganographic protocol that shows how quantum information can be hidden in the noise of a depolarizing channel, using a shared classical secret key between Alice and Bob. We first consider the case when the physical channel is noiseless (i.e., all noise is controlled by Alice), but Eve expects some level of noise; we then extend this to the case where the channel has intrinsic noise (not controlled by Alice and Bob). We calculate the amount of secret key consumed. We then present a quantum steganographic protocol for a general quantum channel, that hides quantum information in the space of typical error syndromes.

We present a mathematical criterion for secrecy based on the diamond norm, and show that, if Alice and Bob have a sufficiently accurate knowledge of Eve's expectations, they can make their communications arbitrarily difficult to detect. In the case where Eve has perfect knowledge of the channel, Alice and Bob can send a finite (but arbitrarily large) amount of hidden information through the channel; if Eve's knowledge of the channel is imprecise, Alice and Bob can communicate at a nonzero asymptotic rate (given an arbitrarily large secret key). We conclude by discussing open questions about achievable rates for arbitrary channels.

II. THE QUANTUM DEPOLARIZING CHANNEL

The quantum analog of the classical binary symmetric channel (BSC) is the depolarizing channel (DC), which is one of the most widely used quantum channel models:

$$\rho \to \mathcal{N}\rho = (1-p)\rho + \frac{p}{3}X\rho X + \frac{p}{3}Y\rho Y + \frac{p}{3}Z\rho Z. \quad (1)$$

That is, each qubit has an equal probability of undergoing an X, Y, or Z error. Applying this channel repeatedly to a qubit will map it eventually to the maximally mixed state $I/2$. We can rewrite this channel in a different but equivalent form:

$$\mathcal{N} = (1 - 4p/3)\mathcal{I} + (4p/3)\mathcal{T}. \quad (2)$$

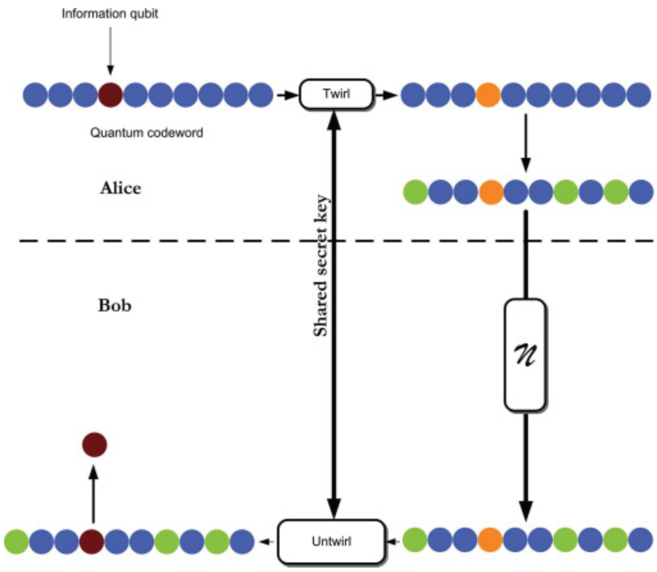

FIG. 2. (Color online) Alice hides her information qubit (solid brown circle) by swapping it in with a qubit of her quantum codeword. She uses her shared secret key with Bob to determine which qubit to swap. She uses the shared key again to twirl the information qubit. She further applies random depolarizing errors to the rest of the qubits of the codeword (shown in green). She sends the codeword through a depolarizing channel to Bob who uses the shared secret to correctly apply the untwirling operation, followed by locating and swapping out Alice's original information qubit.

where $\mathcal{I}\rho = \rho$ and $\mathcal{T}\rho = (1/4)(\rho + X\rho X + Y\rho Y + Z\rho Z)$. The operation $\mathcal{T}$ is twirling: it takes a qubit in any state ρ to the maximally mixed state $I/2$. If we rewrite the channel in this way, instead of applying X, Y, or Z errors with probability $p/3$, we can think of removing the qubit with probability $4p/3$ and replacing it with a maximally mixed state. This picture makes the steganographic protocol more transparent. We will first assume that the actual physical channel between Alice and Bob is noiseless. All the noise that Eve sees is due to deliberate errors that Alice applies to her codewords. We depict this protocol in Fig. 2.

(1) Alice encodes a covertext of k_c qubits into N qubits with an $[[N,k_c]]$ quantum error-correcting code (QECC).

(2) From Eq. (2), the DC would maximally mix Q qubits with probability p_Q where

$$p_Q = \binom{N}{Q}(4p/3)^Q(1-4p/3)^{N-Q}. \tag{3}$$

For large N, Alice can send $M = (4/3)pN(1-\delta)$ stego qubits, where $1 \gg \delta \gg \sqrt{(1-4p/3)/(4pN/3)}$. (The chance of fewer than M errors is negligibly small.)

(3) Using the shared random key (or shared ebits), Alice chooses a random subset of M qubits out of the N, and swaps her M stego qubits for those qubits of the codeword. She also replaces a random number m of qubits outside this subset with maximally mixed qubits, so that the total $Q = M + m$ matches the binomial distribution given in Eq. (3) to high accuracy.

(4) Alice "twirls" her M stego qubits using $2M$ bits of secret key or $2M$ shared ebits. To each qubit she applies one of I, X, Y, or Z chosen at random, so $\rho \to \mathcal{T}\rho$. To Eve, who does not have the key, these qubits appear maximally mixed. (Twirling can be thought of as the quantum equivalent of a one-time pad.)

(5) Alice transmits the codeword to Bob. From the secret key, he knows the correct subset of M qubits and the one-time pad to decode them.

This protocol transmits $(4/3)pN(1-\delta)$ secret qubits from Alice to Bob (Fig. 2). The secrecy follows from this argument: without the key, Eve cannot distinguish a stego qubit from a maximally mixed qubit; and these maximally qubits are distributed exactly as would be expected from the depolarizing channel with error rate p. If the rate p matches Eve's expectations, she will detect nothing suspicious even if she intercepts the codeword and measures its error syndromes.

If the channel contains intrinsic noise, Alice will first have to encode her k_s stego qubits in an $[[M,k_s]]$ QECC, swap those M qubits for a random subset of M qubits in the codeword, and apply the twirling procedure. This twirling does not interfere with the error-correcting power of the QECC if Bob knows the key. The rate of transmission k_s/N will depend on the rate of the QECC used to protect the stego qubits. For a BSC this would be at best $(1-\delta)[1-h(p)]\delta p/(1-2p)$. However, for most quantum channels (including the DC) the achievable rate is not known. Assuming the physical channel is also a DC with error rate p and that Alice emulates a DC with error rate q, the effective channel will appear to Eve like a DC with error rate $p + q(1-4p/3) \equiv p + \delta p$. As long as $p + \delta p$ is sufficiently close to Eve's expectation of the error rate, the communication will remain secret. (We will make this notion of secrecy precise later in the paper.) The rate of communication is $k_s/N \approx (4/3)c\delta p/(1-4p/3)$, where $c = k_s/M$ is the achievable rate of the code for the DC with error rate p.

The secret key is used at two points in these protocols. First, in step (3) Alice chooses a random subset of M qubits out of the N-qubit codeword. There are $C(N,M)$ subsets, so roughly $\log_2 C(N,M)$ bits are needed to choose one. Next, in step (4), $2M$ bits of key are used for twirling. This gives us

$$n_k \approx \log_2 \binom{N}{M} + 2M \tag{4}$$

bits of secret key used. Define the key consumption rate $\mathcal{K} = n_k/N$ to be the number of bits of key consumed per qubit that Alice sends through the channel. We use $M \approx 4qN/3$ and $q \approx \delta p/(1-4p/3)$ to express $\mathcal{K}$ in terms of p, δp, and N (Fig. 3):

$$\mathcal{K} \approx \log_2[(4/\beta)^\beta(1-\beta N)^{\beta-1}], \quad \beta \equiv 4\delta p/(3-4p). \tag{5}$$

Alice can consume fewer bits of key if she has a source that averages to a maximally mixed state—for instance, if Alice first compresses the state $|\phi_s\rangle$ before sending it. This would allow them to bypass the twirling procedure. However, while the secrecy criterion may still be met without twirling, the security criterion would not be: if Eve becomes aware of the message, she may be able to read it without the key.

III. MORE GENERAL CHANNELS

The protocols given above perform well in emulating a depolarizing channel. However, there are far more general

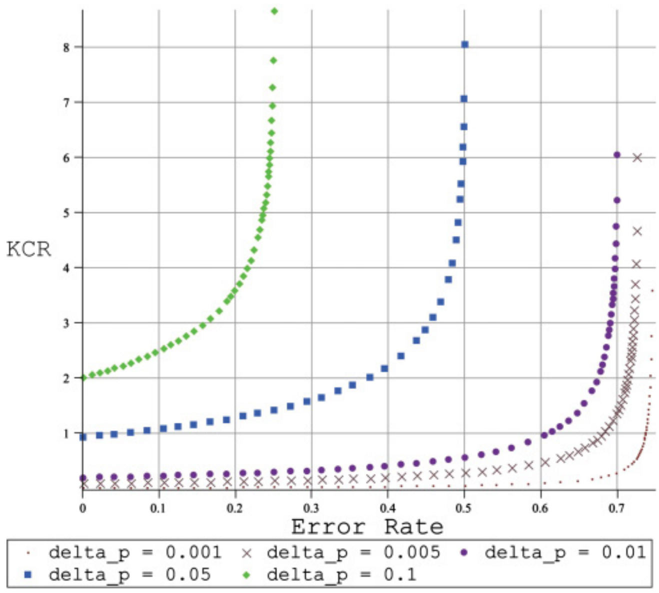

FIG. 3. (Color online) We plot the key consumption rate (KCR) as a function of the error-rate p of the binary-symmetric channel (BSC), with δp being extra uncontrollable noise in the BSC.

channels, and the protocols may not work well, or at all, in other cases. If one has a channel that can be written

$$\rho \to \mathcal{N}\rho = (1 - p_T + p_E)\mathcal{I}\rho + p_T\mathcal{T}\rho + p_E\mathcal{E}\rho, \quad (6)$$

where $\mathcal{E}$ is an arbitrary error operation, one can still use the above protocols to hide approximately $p_T N$ stego bits or qubits, while generating $p_E N$ random errors of type $\mathcal{E}$. But for some channels, p_T may be very small or zero. How should we proceed? Moreover, hiding stego qubits locally as maximally mixed qubits sacrifices some potential information. The location of the error—that is, the choice of the subset holding the errors—could also be used to convey information, potentially increasing the rate and reducing the amount of secret key or shared entanglement required.

A different approach is instead to encode information in the error syndromes. For simplicity, we consider the case when N is large. In this case, it suffices to consider only typical errors. We begin with the case where the physical channel is noise-free.

For large N, almost all (probability $1 - \epsilon$) combinations of errors on the individual qubits will correspond to one of the set of typical errors. There are roughly 2^{sN} of these, and their probabilities p_e are all bounded within a range $2^{-N(s+\delta)} \leqslant p_e \leqslant 2^{-N(s-\delta)}$. The number s is the entropy of the channel on one qubit; for the BSC, $s = h(p) = -p\log_2 p - (1-p)\log_2(1-p)$, and for the DC, $s = -(1-p)\log_2(1-p) - p\log_2(p/3)$. We label the typical error operators $E_0, E_1, \ldots, E_{2^{sN}-1}$, and their corresponding probabilities are p_j. A good choice of QECC for the cover text will be able to correct all these errors. We make the simplifying assumption that the QECC is nondegenerate, so each typical error E_j has a distinct error syndrome labeled s_j.

Ahead of time, Alice and Bob partition the typical errors into C roughly equiprobable sets S_k, so that

$$\sum_{E_j \in S_k} p_j \approx \frac{1}{C}\ \forall\, k. \quad (7)$$

As far as possible, the errors in a given set should be chosen to have roughly equal probabilities. The maximum of C is roughly $C \approx 2^{N(s-\delta)}$, and $k = 0, \ldots, C-1$. We can now present a new quantum steganographic protocol using error syndromes to store information.

(1) Alice prepares k_c qubits of cover text in a state $|\psi_c\rangle$.

(2) Alice's secret message is a string of $\log_2 C \approx N(s-\delta)$ qubits, in a state

$$|\psi_s\rangle = \sum_{k=0}^{C-1} \alpha_k |k\rangle. \quad (8)$$

She "twirls" each qubit of this string, using $2N(s-\delta)$ bits of the secret key or shared ebits, to get a maximally mixed state. To this, she appends $N - k_c - (s-\delta)N$ extra ancilla qubits in the state $|0\rangle$ to make up a total register of $N - k_c$ qubits.

(3) Using the shared secret key, Alice chooses from each set S_k a typical error E_{j_k} with syndrome s_{j_k}. She applies a unitary operator U_S to the register of $N - k_c$ qubits, that maps $U_S\left(|k\rangle \otimes |0\rangle^{\otimes N-k_c-(s-\delta)N}\right) = |s_{j_k}\rangle$. She appends this register to the cover qubits in state $|\psi_c\rangle$, then applies the encoding unitary U_E. Averaging over the secret key, the resulting state will appear to Eve like $\rho \approx \sum_{j=0}^{2^{nS}-1} p_j E_j |\Psi_c\rangle\langle\Psi_c| E_j^\dagger$, which is effectively indistinguishable from the channel being emulated acting on the encoded cover text.

(4) Alice sends this codeword to Bob. If Eve examines its syndrome, she will find a typical error for the channel being emulated.

(5) Bob applies the decoding unitary $U_D = U_E^\dagger$, and then applies $U_S^\dagger$ (which he knows using the shared secret key). He discards the cover text and the last $N - k_c - (s-\delta)N$ ancilla qubits and undoes the twirling operation on the remaining qubits, again using the secret key. If Eve has not measured the qubits, he will have recovered the state encoded by Alice.

This protocol may easily be used to send classical information by using a single basis state rather than a superposition like Eq. (8). The steganographic transmission rate $\mathcal{R}$ is roughly $\mathcal{R} \approx s - \delta \to s$. The rate of transmission s is higher than the rate $4p/3$ of our first protocol. This protocol used $2N(s-\delta)$ bits of secret key (or ebits) for twirling in step (2), and roughly $N\delta$ bits of secret key in choosing representative errors E_{j_k} from each set S_k in step (3). So the key rate is roughly $\mathcal{K} \approx 2s - \delta \to 2s$, which is better than the first protocol in key usage per stego qubit transmitted. Since almost all the key usage goes to the twirling operation, for sources that are maximally mixed on average the rate of key usage can actually go to zero as $N \to \infty$. However, this encoding is much trickier in the case where the channel contains intrinsic noise.

In principle this quantum steganographic protocol can be used when the channel contains noise. The steganographic qubits are first encoded in a QECC to protect them against the noise in the channel. In practice, for many channels this can be difficult—the effects of errors on the space of syndromes look quite different from a usual additive error channel. Also, unlike the depolarizing channel, general channels when composed together may change their type. However, by drawing on codes with suitable properties, the problem of designing steganographic protocols for general channels may be simplified. We analyze the special case of the binary symmetric channel in Appendix B, but the solution for a general channel is a problem for future work.

IV. SECRECY AND SECURITY

What is the standard of security for a stego protocol? There are two obvious considerations. First, if Eve becomes suspicious, can she read the message? At the cost of using one-time pads or twirling, Alice and Bob can prevent this from happening. This is the question of security.

The more important question is: can Alice and Bob avoid arousing Eve's suspicions in the first place? This is the question of secrecy. To do this, the messages that Alice sends must emulate as closely as possible the channel that Eve expects. We can make this condition quantitative. Let $\mathcal{E}_C$ be the channel on N qubits that Eve expects, and let $\mathcal{E}_S$ be the effective channel that Alice and Bob produce with their steganographic protocol. Then the protocol is secure if $\mathcal{E}_S$ is ϵ-close to $\mathcal{E}_C$ in the diamond norm $\|\mathcal{E}_S - \mathcal{E}_C\|_\diamond \leqslant \epsilon$ for some small $\epsilon > 0$. The diamond norm is directly related to the probability for Eve to distinguish $\mathcal{E}_C$ from $\mathcal{E}_S$ under ideal circumstances (i.e., when she controls both inputs and outputs), and so puts an upper bound on her ability to distinguish them in practice.

For a simple example, the difference between two DCs applied to N qubits has norm

$$\left\|\mathcal{N}_r^{\otimes N} - \mathcal{N}_p^{\otimes N}\right\|_\diamond = \sum_{j=0}^{N} \binom{N}{j} |r^j (1-r)^{N-j} - p^j (1-p)^{N-j}|, \qquad (9)$$

where p is the error rate of the channel Eve expects and $r = p + \delta p$ is the error rate of the steganographic channel that emulates Eve's expected channel. If we make $\delta p < \epsilon\sqrt{p(1-p)/N}$ then we can make this norm as small as we like, while communicating $O(\delta p N) = O(\epsilon\sqrt{N})$ secret qubits. This indicates that even if Eve has exact knowledge of the channel, Alice and Bob can, in principle, send an arbitrarily large (but finite) amount of information without arousing Eve's suspicion, by choosing a sufficiently small δp and large N. If Eve's knowledge of the channel is imperfect, Alice and Bob can do even better, communicating steganographic information at a nonzero rate.

If Eve is constantly monitoring the channel over a long period of time, and if she has exact knowledge of the channel then she will eventually learn that Alice and Bob are communicating with each other steganographically. Moreover, with constant measurement Eve can disrupt the superpositions of the steganographic qubits and prevent any quantum information from ever reaching Bob, effectively flooding the quantum channel with noise (though classical communication will still be possible). We calculate the diamond norm for the BSC and DC cases in Appendix A.

If Alice and Bob share a secret, random key, they can use the steganographic encodings described in the paper. Shared entanglement (ebits) can act as a resource in the same way—by measuring the two halves of a maximally entangled pair of qubits $(|00\rangle + |11\rangle)/\sqrt{2}$ Alice and Bob can generate a shared secret bit.

However, the use of ebits does open up an additional possibility beyond what can be done with a classical key. Instead of sending quantum information through the channel, Alice can instead teleport qubits to Bob. Teleportation consumes one ebit and requires the transmission of two classical bits for each qubit teleported. These classical bits can be sent through the channel steganographically. Because these bits are perfectly random, no one-time pad or twirling is needed. And because they are purely classical information, they are not disrupted if Eve chooses to measure the error syndromes, as a general quantum state would be. In this sense, quantum steganography with shared ebits is more powerful than quantum steganography with a shared classical key.

V. CONCLUSIONS

The problem of quantum steganography is to send hidden quantum information through a quantum channel in such a way that an eavesdropper with the power to monitor the channel will be unaware that the secret communication has taken place. Our approach to this problem is to disguise the information as channel errors on a codeword containing "innocent" information. We have presented two different protocols to achieve this goal, given a quantitative measure of their secrecy, and calculated rates for communication and secret key usage. The second protocol, in which information is encoded in the space of error syndromes, is suitable for general channels, and achieves higher communication rates with lower use of the secret key. However, for this approach it is unclear what the best encoding is when the channel contains intrinsic noise. This is the subject of ongoing research. Quantum steganography represents a new type of cryptographic protocol for quantum information, and opens many fascinating new questions.

ACKNOWLEDGMENTS

BAS and TAB would like to thank Daniel Gottesman, Patrick Hayden, Debbie Leung, John Preskill, Stephanie Wehner, and Mark Wilde for their useful comments and suggestions at various stages of this work. This work was funded in part by NSF Grant No. CCF-0448658. TAB also acknowledges the hospitality and support of the Kavli Institute for Theoretical Physics.

APPENDIX A: DIAMOND NORM

In defining the diamond norm we follow the conventions from Nielsen and Chuang [8] and John Watrous's lecture notes [9]. The diamond norm give a measure of how "close" or similar two channels can be when they transform an arbitrary density matrix from one Hilbert space to another. We use this as a quality measure for the "innocence" of the quantum message from Alice to Bob: if Eve cannot distinguish a channel containing the steganographic message from the channel that she expects, the stego channel satisfies the secrecy criterion.

Let $\mathcal{N}$ be some arbitrary super-operator and let $\mathcal{N}: L(\mathcal{V}) \to L(\mathcal{W})$, where $L(.)$ is a space of linear operators on the Hilbert spaces $\mathcal{V}$ and $\mathcal{W}$. Then one can define the diamond norm of $\mathcal{N}$ as

$$\|\mathcal{N}\|_\diamond \equiv \|I_{L(\mathcal{V})} \otimes \mathcal{N}\|_{\mathrm{tr}}, \tag{A1}$$

where $\|\mathcal{N}\|_{\mathrm{tr}}$ is defined as

$$\|\mathcal{N}\|_{tr} \equiv \max\{\|\mathcal{N}(O)\|_{\mathrm{tr}} : O \in L(\mathcal{V}), \|O\|_{\mathrm{tr}} = 1\}. \tag{A2}$$

The maximization in Eq. (A2) is over all density matrices. When the Hilbert space is infinite dimensional we take the supremum of the set defined in (S2).

1. Binary symmetric channel

Let $0 < p < 1/2$ be the rate at which Alice flips the qubits of her codeword. Let $r \equiv p + \delta p$ be the rate at which the BSC flips qubits, where δp is some additional noise which is not under the control of either Alice or Bob. We assume that $0 < p < r < 1/2$ because, at $p = 1/2$, the channel has zero capacity to send information and $p > 1/2$ means that most qubits are being flipped, which is unnatural for this channel. For a single qubit ($N = 1$), let $\mathcal{N}_p$ be the BSC that Alice applies to an arbitrary single-qubit density operator ρ:

$$\mathcal{N}_p\rho = (1-p)\rho + pX\rho X, \tag{A3}$$

and let $\mathcal{N}_r$ be the actual BSC:

$$\mathcal{N}_r\rho = (1-r)\rho + rX\rho X. \tag{A4}$$

We can now express the difference of the two channels as

$$(\mathcal{N}_r - \mathcal{N}_p)\rho = (p-r)\rho + (r-p)X\rho X. \tag{A5}$$

We can express the diamond norm of the difference of the channels $\mathcal{N}_p$ and $\mathcal{N}_r$ as

$$\|\mathcal{N}_r - \mathcal{N}_p\|_\diamond = \max_\rho \|[I \otimes (\mathcal{N}_r - \mathcal{N}_p)]\rho\|_{\mathrm{tr}} \tag{A6}$$

$$= (r-p)\max_\rho \|(I \otimes I)\rho(I \otimes I) - (I \otimes X)\rho(I \otimes X)\|_{\mathrm{tr}}. \tag{A7}$$

When we substitute $\rho = \psi \otimes |0\rangle\langle 0|$ (ψ is some arbitrary density operator) in the above equation we achieve the maximum.

$$\begin{aligned}\|\mathcal{N}_r - \mathcal{N}_p\|_\diamond &= (r-p)\|\psi \otimes |0\rangle\langle 0| - \psi \otimes |1\rangle\langle 1|\|_{\mathrm{tr}} \\ &\leqslant (r-p)\|\psi\|_{\mathrm{tr}}\||0\rangle\langle 0|\|_{\mathrm{tr}} + \|\psi\|_{\mathrm{tr}}\||1\rangle\langle 1|\|_{\mathrm{tr}} \\ &= (r-p)(1+1) = 2(r-p) \\ &= 2(p+\delta p - p) = 2\delta p.\end{aligned} \tag{A8}$$

In Eq. (A8), we use the triangle inequality and we use the fact that, for any two linear operators A and B, the trace norm of their tensor product is equal to the product of their trace norms, (i.e., $\|A \otimes B\|_{\mathrm{tr}} = \|A\|_{\mathrm{tr}}\|B\|_{\mathrm{tr}}$). We would like an expression for the optimal probability to correctly distinguish two channels.

$$P_{\mathrm{opt}} = \frac{1}{2} + \frac{1}{4}\left\|\mathcal{N}_r - \mathcal{N}_p\right\|_\diamond. \tag{A9}$$

So for a single qubit use

$$P_{\mathrm{opt}} = \frac{1}{2}(1+\delta p). \tag{A10}$$

For the case where we have two qubits, we can write Alice's BSC as

$$\begin{aligned}(\mathcal{N}_p \otimes \mathcal{N}_p)\rho = (1-p)^2\rho + p(1-p)X_1\rho X_1 \\ + p(1-p)X_2\rho X_2 + p^2 X_1X_2\rho X_1X_2,\end{aligned}$$

where $X_1 \equiv X \otimes I$, $X_2 \equiv I \otimes X$, and $X_1X_2 \equiv X \otimes X$. We can similarly calculate $\mathcal{N}_1 \otimes \mathcal{N}_1$. We can now write the difference between the two channels as

$$\begin{aligned}&(\mathcal{N}_r \otimes \mathcal{N}_r - \mathcal{N}_p \otimes \mathcal{N}_p)\rho \\ &\quad = (r^2 - 2r + 2p - p^2) \\ &\qquad + (r - r^2 - p + p^2)(X_1\rho X_1 + X_2\rho X_2) \\ &\qquad + (r^2 - p^2)X_1X_2\rho X_1X_2.\end{aligned} \tag{A11}$$

The diamond norm of the difference between two BSC on two qubits can be expressed as

$$\begin{aligned}&\|\mathcal{N}_r \otimes \mathcal{N}_r - \mathcal{N}_p \otimes \mathcal{N}_p\|_\diamond \\ &\quad = \max_\rho \|[I \otimes (\mathcal{N}_r \otimes \mathcal{N}_r - \mathcal{N}_p \otimes \mathcal{N}_p)]\rho\|_{\mathrm{tr}}.\end{aligned} \tag{A12}$$

We use a similar construction from the single-qubit case to maximize the right side of Eq. (A12). Letting $\rho = \psi \otimes |00\rangle\langle 00|$ in Eq. (A12), we get

$$\begin{aligned}&\|\mathcal{N}_r \otimes \mathcal{N}_r - \mathcal{N}_p \otimes \mathcal{N}_p\|_\diamond \\ &\quad = |(1-r)^2 - (1-p)^2| \\ &\qquad + 2|r(1-r) - p(1-p)| \\ &\qquad + |r^2 - p^2|.\end{aligned} \tag{A13}$$

Given our constraints that $0 < p < r < 1/2$, the first term on the right side of Eq. (A13) is negative while the second and third terms are positive. This give us

$$\begin{aligned}\|\mathcal{N}_r \otimes \mathcal{N}_r - \mathcal{N}_p \otimes \mathcal{N}_p\|_\diamond &= 2(r-p)(2-r-p) \\ &= 2\delta p(2 - 2p - 2\delta p).\end{aligned}$$

So in the double-qubit case P_{opt} is

$$P_{\mathrm{opt}} = \frac{1}{2}[1 + \delta p(2 - 2p - 2r)]. \tag{A14}$$

If we observe Eq. (A13) carefully we find that the terms are distributed binomially. For the case where we have N qubits, we can use $\rho = \psi \otimes |00\cdots 0\rangle\langle 00\cdots 0|$ to maximize the diamond norm for N uses of BSC to get

$$\begin{aligned}\left\|\mathcal{N}_r^{\otimes N} - \mathcal{N}_p^{\otimes N}\right\|_\diamond = \sum_{j=0}^{N}\binom{N}{j}|r^j(1-r)^{N-j} \\ - p^j(1-p)^{N-j}|.\end{aligned} \tag{A15}$$

2. Depolarizing Channel

The calculation of the diamond norm of the difference between N uses of two depolarizing channels (DC) is similar to the calculation of BSC that we performed in the previous section. The expression for the channel is

$$\mathcal{N}_p\rho = (1-p)\rho + (p/3)(X\rho X + Y\rho Y + Z\rho Z). \tag{A16}$$

Eve sees a channel with a somewhat higher rate $r = p + \delta p$. As in the BSC case we assume that $0 < p < r < 1/2$. For the case $N=2$ the difference between the two depolarizing channels is

$$\begin{aligned}(\mathcal{N}_r \otimes \mathcal{N}_r - \mathcal{N}_p \otimes \mathcal{N}_p)\rho &= [(1-r)^2 - (1-p)^2]\rho \\ &\quad + [(1-r)(r/3) - (1-p)(p/3)] \\ &\quad \times (X_1 \rho X_1 + \cdots + Z_2 \rho Z_2) \\ &\quad + [(r/3)^2 - (p/3)^2] \\ &\quad \times (X_1 X_2 \rho X_1 X_2 + \cdots + Z_1 Z_2 \rho Z_1 Z_2).\end{aligned}$$

The density matrix that maximizes the trace norm is $\rho = \psi \otimes |\Phi^+\rangle\langle\Phi^+|$, where $|\Phi^+\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$, and ψ is some arbitrary single-qubit density operator.

$$\begin{aligned}\|\mathcal{N}_r \otimes \mathcal{N}_r - \mathcal{N}_p \otimes \mathcal{N}_p\|_\diamond &= |(1-r)^2 - (1-p)^2| \\ &\quad + 6|(1-r)(r/3) - (1-p)(p/3)| \\ &\quad \times 9|(r/3)^2 - (p/3)^2| \\ &\quad + |(1-r)^2 - (1-p)^2| \\ &\quad + 2|(1-r)r - (1-p)p| + |r^2 - p^2|. \end{aligned} \tag{A17}$$

After evaluating the absolute value terms, we get

$$\begin{aligned}\|\mathcal{N}_r \otimes \mathcal{N}_r - \mathcal{N}_p \otimes \mathcal{N}_p\|_\diamond &= 2(r-p)(2-r-p) \\ &= 2\delta p(2-2p-\delta p).\end{aligned}$$

So,

$$P_{\rm opt} = \frac{1}{2} + \frac{1}{2}\delta p(2-2p-\delta p). \tag{A18}$$

For the general case of N uses of the depolarizing channel, we may write the diamond norm as

$$\begin{aligned}&\left\|\mathcal{N}_r^{\otimes N} - \mathcal{N}_p^{\otimes N}\right\|_\diamond \\ &= \sum_{j=0}^{N} \binom{N}{j} |r^j(1-r)^{N-j} - p^j(1-p)^{N-j}|,\end{aligned} \tag{A19}$$

which is exactly the same expression as for the BSC.

APPENDIX B: PROPERTIES OF PROTOCOL 2

1. Achievable rate

We will work out the simplest example—the BSC in the case where the physical channel is noise-free. The errors in the codewords that Alice sends to Bob are binomially distributed. Let pN be the mean of this distribution and let the variance be $pN\delta$, where $0 < \delta \ll 1$. Here, N is the length of each of codeword. Let

$$p_k = \binom{N}{k} p^k (1-p)^{N-k} \tag{B1}$$

be the errors that Alice applies to her codewords. For each k from $Np(1-\delta)$ to $Np(1+\delta)$ choose C_k strings of weight k. Let

$$C = \sum_{k=Np(1-\delta)}^{Np(1+\delta)} C_k. \tag{B2}$$

Let these sets of strings be called S_k, and

$$S = \cup_k S_k. \tag{B3}$$

So the total number of strings in the set S is C. Define the probability $q \equiv 1/C$. Then we want to satisfy $qC_k = C_k/C = p_k$. Clearly we must have $\binom{C_k}{k} \leqslant N$ for all k. This implies that

$$\begin{aligned}C_k p^k (1-p)^{N-k} &\leqslant \binom{N}{k} p^k (1-p)^{N-k}, \\ \Rightarrow C_k p^k (1-p)^{N-k} &\leqslant C_k q, \\ \Rightarrow p^k (1-p)^{N-k} &\leqslant q.\end{aligned}$$

We want C to be as large as possible, which means we want q to be as small as possible. This constraint then gives us

$$\begin{aligned}q &= p^{Np(1-\delta)}(1-p)^{N(1-p+p\delta)}, \\ \Rightarrow C &= 1/q, \\ \Rightarrow C &= p^{-Np(1-\delta)}(1-p)^{-N(1-p+p\delta)}.\end{aligned}$$

The number of bits that Alice can send is, therefore,

$$\begin{aligned}M &= \log_2 C \\ &= N\{-p\log_2 p - (1-p)\log_2(1-p). \\ &\qquad + \delta[p\log_2 p - p\log_2(1-p)]\} \\ &= N\{h(p) - p\delta \log_2[(1-p)/p]\}.\end{aligned}$$

So with this encoding Alice can send almost $Nh(p)$ bits.

2. Diamond norm

Again we consider the simplest case of the BSC. Let N be sufficiently large so that the total probability of the typical errors is $> 1-\epsilon$, and these typical errors have weight k in the range $Np(1-\delta) \leqslant k \leqslant Np(1+\delta)$. We divide up all errors of weight k into C_k partitions containing

$$n_k \approx \frac{\binom{N}{k}}{C_k} \approx \left(\frac{1-p}{p}\right)^{k-Np(1-\delta)}$$

errors each. Within each set the errors are all equally likely to be chosen. However, because the number of errors is unlikely to divide exactly evenly into C_k sets, the probabilities q_k of an error of weight k will be slightly different from the probability $p_k = p^k(1-p)^{N-k}$ of the binomial distribution. We can put a (not-very-tight) bound on this difference:

$$|q_k - p_k| < \frac{p_k}{\left(\frac{1-p}{p}\right)^{k-Np(1-\delta)} - 1} < \frac{1-p}{1-2p} p^{2k}(1-p)^{N-2k}. \tag{B4}$$

Plugging this into the expression for the diamond norm, we get

$$\begin{aligned}\left\|\mathcal{N}_p^{\otimes N} - \mathcal{N}_{\rm enc}\right\|_\diamond &< \epsilon + \sum_{k=Np(1-\delta)+1}^{Np(1+\delta)} \binom{N}{k} |p_k - q_k| \\ &< \epsilon + \left(\frac{1-p}{1-2p}\right)\left(\frac{p}{1-p}\right)^{Np(1-\delta)} \\ &\quad \times \left(\frac{1-2p+2p^2}{1-p}\right)^N,\end{aligned}$$

which is exponentially small in N.

3. Error correction with a noisy channel

Since errors can act in a complicated manner on the space of syndromes, it is not entirely clear what the optimal encoding is even for a simple channel. Here, we present one encoding for the BSC that gives an achievable rate in the limit of large N, but it is quite likely that higher rates are possible.

In the noiseless case, it is possible to use the $C(N,M)$ strings of weight M as a code—each string represents one possible weight-M error. If we then apply a BSC with probability p, on average Np bits would be flipped. If $Np \ll M$ then one can keep only a subset of the weight-M strings, separated by a distance $> 2Np$.

This encoding quickly becomes inefficient as p gets larger. Using the shared secret key, Alice can instead choose only a subset of the N bits to hold the codewords. If this subset includes N' bits, then the errors on the remaining $N - N'$ bits are irrelevant and do not need to be corrected. The limit of this would be similar to encoding 1 in the paper, where $N' \approx 2M$.

Let $N' = qN$ for some $0 < q \leqslant 1$. The number of strings of weight M is $C(qN,M)$, and there will be an average number of bit flips pqN on the relevant portion of the codeword. Keep a subset of these codewords separated by distance $2pqN$. Decoding is done by finding the closest codeword to the output string.

As $N,M \to \infty$ then the number of codewords will go like

$$C(N,M,p,q) \sim \frac{\binom{qN}{M}}{\binom{qN}{pqN}}.$$

The number of bits will be $\log_2 C(N,p,q)$.

Since q is a parameter we can choose freely, we choose it to maximize the rate $\mathcal{R}(N,M,p,q) \equiv (1/N)\log_2 C(N,M,p,q)$. Using the Stirling approximation, differentiating with respect to q, and setting the result equal to 0, we can solve for q:

$$q = \frac{M}{N}\left(\frac{2^{h(p)}}{2^{h(p)} - 1}\right).$$

We can then plug this back into the formula for $\mathcal{R}$. If the physical channel has error rate p and Alice is attempting to emulate a channel with error rate $p + \delta p$, then $M = N\delta p/(1-2p)$. This gives us the following expression for the rate:

$$\mathcal{R}(p,\delta p) = -\frac{\delta p}{1-2p}\log_2(2^{h(p)} - 1).$$

We can compare this to the rate from encoding 1, which for the BSC is $2\delta p[1 - h(p)]/(1-2p)$. It is not hard to see that $\mathcal{R}(p,\delta p)$ above approaches this rate as $p \to 1/2$ (and both rates go to zero), but as $p \to 0$ this encoding does considerably better than encoding 1. It is quite likely, however, that there may be even more efficient encodings.

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

[2] J. Trithemius, *Steganographia* (First printed edition, Frankfurt, 1606).

[3] G. J. Simmons, in *Advances in Cryptology: Proceedings of Crypto 83*, edited by D. Chaum (Plenum Press, New York, 1984), pp. 51-67.

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

[5] M. Curty and D. J. Santos, in *2nd Bielefeld Workshop on Quantum Information and Complexity, Bielefeld, Germany, 2004*.

[6] S. Natori, Quantum Computation and Information **102**, 235 (2006).

[7] K. Martin, Lect. Notes Comput. Sci. **4567**, 32 (2007).

[8] I. Chuang and M. Nielsen, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).

[9] J. Watrous, Lecture notes available at [http://www.cs.uwaterloo.ca/∼watrous/quant-info] (2008).