Hiding messages in quantum data

Julio Gea-Banacloche

Citation: Journal of Mathematical Physics **43**, 4531 (2002); doi: 10.1063/1.1495073
View online: https://doi.org/10.1063/1.1495073
View Table of Contents: http://aip.scitation.org/toc/jmp/43/9
Published by the American Institute of Physics

Hiding messages in quantum data

Julio Gea-Banacloche[a)]
Department of Physics, University of Arkansas, Fayetteville, Arkansas 72701

(Received 15 January 2002; accepted for publication 16 May 2002)

A method is proposed to hide messages in arbitrary quantum data files. The messages may act as ''watermarks,'' to secure the authenticity and/or integrity of the data. With the help of classical secret keys, they can be made unreadable by other parties and to reveal whether thay have been tampered with. The basic idea is to encode the data using a quantum error-correcting code and hide the message as (correctible) errors, deliberately inserted, which can be read out from the error syndrome. Also discussed briefly is a ''reverse encoding,'' which would involve putting the actual data in the error syndrome, and letting the encoded qubit itself carry the message. © *2002 American Institute of Physics.*
[DOI: 10.1063/1.1495073]

I. INTRODUCTION

Steganography is a branch of cryptography concerned with embedding ''invisible'' messages in data files. The message, which is revealed by some appropriate decoding operation, may contain information regarding the owner of the file or its date of creation, for instance. Such techniques have become of particular interest in recent years because of the proliferation of means available to copy, legally or illegally, all sorts of data, such as images, audio, or video files.[1]

Although quantum information processing is still, in practice, a long way from the day where sizable files of ''quantum data'' will be moved from one location to another, it may already be of interest to begin to explore the possibilities inherent in a quantum sort of steganography, and the ways in which this would differ from its classical counterpart. This article is intended to serve as a (small) first step in this direction.

It should be clear from the outset, of course, that copyright protection in its most literal sense could never be an issue for quantum information, since, by the celebrated no-cloning theorem, it is inherently impossible to copy. Nonetheless, there are other useful purposes that could be served by a hidden message embedded in a quantum data file. It could function as a ''watermark,'' for instance, allowing one to identify the file's owner or creator, either as protection against theft or reassurance to the party receiving the data that they come, in fact, from the right source. Also, as will be shown below, the watermark could be embedded in such a way as to provide the receiving party with information that the file has been corrupted, either by errors upon transmission or by tampering by a third party. Another potential use might be in a distributed quantum computing environment, where packets of information are processed at some location and then sent to other processors: a message could be embedded in the data to tell the receiving processor what it is supposed to do with it. Again, corruption of the data could be detected at the receiving end by these means.

A large component of all of the above is, clearly, the question of data authentication, in the broadest sense of making sure both that the data have come from the right source and that they have not been tampered with or otherwise corrupted. A number of ideas have recently been put forth regarding this general issue. Buhrman *et al.*[2] have studied ways to associate a quantum fingerprint with a classical data string; a variation on this idea, due to Gottesman and Chuang,[3] would allow one to attach a quantum signature to a classical message, which would serve to

[a)]Electronic mail: jgeabana@uark.edu

certify its authenticity. The general question of authenticating a single classical bit using a single qubit has been studied by Curty and Santos.[4] Moving on to the problem of authenticating *quantum* data, Curty *et al.*[5] have shown that it is impossible to authenticate a qubit with a one-bit (or one-qubit) key (in all authentication schemes, the sender and receiver must share a key in advance, in order to read out the authentication information). Nonetheless, if longer keys are allowed, schemes to authenticate arbitrary quantum data exist; for instance, in Ref. 6, Leung has proposed a scheme to authenticate n qubits given an authenticated two-way classical channel, $2r$ bits of classical key, and an extra $2r$ qubits of quantum communication. In this scheme, forging succeeds with probability no better than 2^{-r} and the fidelity of an accepted message with respect to the original is of the order of $1-O(2^{-r})$. Finally, Barnum *et al.*[7] have also recently reported on a powerful general method of quantum data authentication, whose details should soon be made widely available.

In Leung's scheme, the $2r$ extra qubits are simply appended to the message, as a sort of authentication tag. The tag is uniquely related to the message, so that a would-be forger cannot simply remove it and attach it to a false message to make it pass for an authentic one. Nonetheless, the tag could still be simply removed, leaving the adversary in possession of the data, with no indication as to its provenance. In this respect this scheme differs from a classical steganography situation, where (part of) the idea is that the identification tag could not be deleted without damaging the data.

In what follows I wish to present a simple authentication scheme which accomplishes this last purpose, and has other potential advantages (such as the fact that it is based on a conventional quantum error correction code, and so, under some circumstances, it may function as such, thus helping to protect the integrity of the data). Perhaps it should be acknowledged from the outset that, even though the starting point draws some inspiration from classical steganography, the final result may not look much like it. The main guiding principle has been to arrange for the data and the message (or ''tag'') to be inextricably linked, in such a way that somebody who does not have the key could not alter the one without altering the other.

The approach is heuristic throughout; no attempt has been made to provide formal proofs of security or optimize the resources needed. Nonetheless, it is interesting that this simple-minded approach leads in a natural way to a construction which ends up having some similarities to the much more formal approach of Ref. 7. Hence, it is hoped that the present work may provide some insight into some of the issues involved in quantum data authentication.

II. EMBEDDING A MESSAGE IN A QUANTUM DATA FILE ENCODED WITH AN ERROR-CORRECTING CODE

A. Basic concept

One conventional approach in clasical steganography is to write the hidden message in the data file by replacing some of the original data's bits, according to a certain pattern or key. Such an approach, however, could not work, in general, with quantum data, which are often in coherent superpositions for which the actual state of an individual qubit may simply not be defined, due to entanglement with other qubits. Arbitrarily setting the state of a random qubit to a specific value would not only destroy that qubit's entanglement, it would also collapse the state of any other qubits entangled with it.

This means that one needs to add qubits to the original file, perhaps interspersing them among the original data qubits, and write the message in them. Then one may observe that a systematic method to add qubits to a quantum data file, in such a way that one could, at the same time, introduce ''errors'' in it harmlessly, is provided by QECCs.[8] Hence the simplest approach to ''quantum steganography'' might be as follows: encode the original data file using a suitable QECC; hide the message as ''errors'' in the encoded qubits; and read it out in the *error syndrome*.

As an example of this approach, consider the following simple three-qubit code, which protects one logical qubit against a single bit flip error (of any of the three physical qubits):

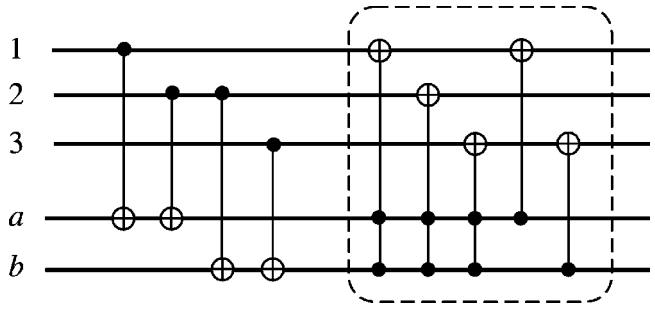

FIG. 1. A circuit to read out the error syndrome for the 3-qubit repetition code. The dashed box shows how the error can be corrected without actually having to measure the state of the ancilla qubits.

$$|0_L\rangle = |000\rangle,$$
$$|1_L\rangle = |111\rangle. \tag{1}$$

The state $|\psi\rangle$ of any of the original data qubits could be encoded as $|\psi\rangle = \alpha|0_L\rangle + \beta|1_L\rangle$. Then, for each such logical qubit, a message of up to four classical bits could be encoded "in parallel," so to speak, on the same three physical qubits, by acting on $|\psi\rangle$ with either the identity or one of the three Pauli operators σ_{ix} (with $i=1, 2, 3$). This message could be read off, for instance, by the error-diagnosing circuit shown on the left half of Fig. 1, which employs two ancillary qubits, a and b. If the state of these ancillary qubits is read out after the four CNOT operations have been carried out, it is easily checked that each of the four possible operations on $|\psi\rangle$ maps to a different joint state of a and b, as follows:

$$1 \rightarrow |00\rangle_{ab},$$
$$\sigma_{1x} \rightarrow |10\rangle_{ab},$$
$$\sigma_{2x} \rightarrow |11\rangle_{ab},$$
$$\sigma_{3x} \rightarrow |01\rangle_{ab}. \tag{2}$$

Thus, all four possible bit values of the two-bit classical mesage can be read and identified, without actually changing the logical qubit's state in any way.

Of course, one problem that arises immediately is that an adversary who knew that the underlying code is the three-qubit, repetition code (1), and wanted to erase the message, could easily do so simply by applying error correction to the encoded qubit! There is, however, a simple way to prevent this from happening, namely, to change the code randomly, from one logical qubit to the next, according to a secret key (shared by the sender and the receiver). For practical purposes, it would be sufficient, for example, to stay with a three-qubit code but encode the message alternatively in bit-flip errors and phase errors. Protection against phase errors is provided by a code like

$$|\overline{0}_L\rangle = H^{\otimes 3}|0_L\rangle = \frac{1}{2^{3/2}}(|0\rangle+|1\rangle)(|0\rangle+|1\rangle)(|0\rangle+|1\rangle),$$
$$|\overline{1}_L\rangle = H^{\otimes 3}|1_L\rangle = \frac{1}{2^{3/2}}(|0\rangle-|1\rangle)(|0\rangle-|1\rangle)(|0\rangle-|1\rangle). \tag{3}$$

(Here, H denotes a Hadamard transform.) Now, a "0" in the classical key could mean that the original qubit is to be encoded as $|\psi\rangle = \alpha|0_L\rangle + \beta|1_L\rangle$, with $|0_L\rangle$ and $|1_L\rangle$ given by (1), and the

message hidden in the error syndrome as any of the four operators identity+one of the σ_{ix}; whereas a "1" in the classical key would mean that the encoding for the original qubit uses the basis $|\overline{0_L}\rangle$ and $|\overline{1_L}\rangle$ given by (3) above, and the error syndrome is encoded using the σ_{iz} operators.

If the eavesdropper does not realize that the message is now encoded as a phase error, he not only cannot read it, but, moreover, he also cannot erase it properly. Suppose that bit-flip error correction is applied to a qubit $\alpha|\overline{0_L}\rangle+\beta|\overline{1_L}\rangle$ originally encoded in the basis (3). The result will be a superposition of the basis states (1); specifically, $\frac{1}{4}$ of the time it will be $\alpha|0_L\rangle+\beta|1_L\rangle$, and $\frac{3}{4}$ of the time it will be $\alpha|1_L\rangle+\beta|0_L\rangle$.[9] Thus, attempts to erase (or, in general, alter) the message result in changes to the data. Needless to say, the message (in the $|\overline{0_L}\rangle,|\overline{1_L}\rangle$ basis) also is changed as a result of this intervention, so if the file is eventually inspected by somebody who knows both the key and the message, he should be able to tell that it has been tampered with.

It is interesting to note that for such a user, who knows both the key and the message, the "errors" deliberately introduced in the encoded qubits do not affect the ability of the code to function as a QECC and protect the data against accidental (additional) errors of the type naturally corrected by the code [e.g., bit errors for the code (1), phase flips for the code (3)]. This can be formally seen as follows: suppose the original (encoded) state is $|\psi\rangle$, and we put an error $\sigma_{i\alpha}$ on it (where i is the qubit index and $\alpha=x,y$ or z). Then suppose another (unknown) error $\sigma_{j\beta}$ occurs; the resulting state is therefore $\sigma_{j\beta}\sigma_{i\alpha}|\psi\rangle$. Then, in order to diagnose properly the error $\sigma_{j\beta}$, all we need to do is apply $\sigma_{i\alpha}$ again to the state, followed by ordinary error correction, since $\sigma_{i\alpha}\sigma_{j\beta}\sigma_{i\alpha}|\psi\rangle=\pm\sigma_{j\beta}|\psi\rangle$. The minus sign only occurs if $i=j$, $\alpha\neq\beta$, and, as an overall sign, it is irrelevant. Therefore, any error that the code could initially diagnose can still be diagnosed correctly. This property alone may make this scheme attractive: if the data are going to be encoded for protection against errors anyway, one might as well take advantage of this to hide a signature in the error syndrome, essentially at no extra cost.

If larger codes are used (to correct for more errors), one can still use essentially the same strategy to protect the message, namely, encode the logical qubits in either the original $|0_L\rangle,|1_L\rangle$, or the conjugate, $|\overline{0_L}\rangle,|\overline{1_L}\rangle$, basis (connected by a Hadamard transform) according to the value of a secret key. Atempts to do error correction on the encoded qubit by somebody who does not know the key will result in the state being projected onto the wrong basis half of the time.

From an authentication perspective, however, one may have to wonder about the reverse possibility: could somebody who does not know the key change the data without changing the message? It turns out that this is, indeed, possible if the message is classical, as will be shown in the next subsection.

B. Inserting a quantum message

Consider again the example (1), as illustrated in Fig. 1. Since the encoding and decoding are linear operations, a coherent superposition of errors acting on the encoded data qubit $|\psi\rangle$ will, in fact, yield a coherent superposition of the ancilla qubits; so the "message" does not have to be classical information, it could be quantum information as well. In other words, the operation

$$(\gamma+\delta\sigma_{1x}+\eta\sigma_{2x}+\epsilon\sigma_{3x})(\alpha|0_L\rangle+\beta|1_L\rangle) \tag{4}$$

actually encodes three qubits worth of information in three physical qubits. The one-qubit state $\alpha|0\rangle+\beta|1\rangle$ is in the "data" that one recovers after error correction, whereas the two-qubit state

$$\gamma|00\rangle+\delta|10\rangle+\eta|11\rangle+\epsilon|01\rangle \tag{5}$$

is the state of the ancilla qubits when the error syndrome is extracted as in Fig. 1. Note that the ancilla qubits will only be in a coherent superposition after error correction has been applied to qubits 1, 2, and 3 (otherwise they are still entangled with them), and only if this is done coherently, i.e., without measurements; this can always be achieved, as indicated, e.g., by the circuit in the

dashed box in Fig. 1. Incidentally, since CNOT gates are their own inverses, the dashed box in Fig. 1 also shows how the encoding (4) could be applied to the qubits 1, 2, 3, if one starts with the qubits a and b in the state (5).

Now consider the possibility mentioned at the end of the previous subsection, that an adversary might attempt to change the data while leaving the message intact. For instance, suppose that he attempts to flip an encoded qubit, by applying the operator $\sigma_{1x}\sigma_{2x}\sigma_{3x}$ to it. If the qubit is encoded in the basis (1), as in Eq. (4), this will indeed change it to $\alpha|1_L\rangle+\beta|0_L\rangle$, while leaving the error syndrome intact, since all the σ_{ix} commute with each other. On the other hand, if the qubit is initially encoded in the conjugate basis $|\overline{0_L}\rangle,|\overline{1_L}\rangle$ as $|\psi\rangle=(\gamma+\delta\sigma_{1z}+\eta\sigma_{2z}+\epsilon\sigma_{3z})\times(\alpha|\overline{0_L}\rangle+\beta|\overline{1_L}\rangle)$, one finds

$$\sigma_{1x}\sigma_{2x}\sigma_{3x}|\psi\rangle=(\gamma-\delta\sigma_{1z}-\eta\sigma_{2z}-\epsilon\sigma_{3z})(\alpha|\overline{0_L}\rangle-\beta|\overline{1_L}\rangle). \tag{6}$$

So, in this basis, the data qubit is also changed (it acquires a ''phase flip''), but now also the syndrome is changed, provided γ and at least one of δ, γ or ϵ are nonzero. That is, in order to provide protection against this type of attack, the ''message'' encoded in the error syndrome must be a quantum superposition state: it must be quantum, not classical, information.

C. Reversing the "data" and "message" roles

The construction in the previous subsection suggests the possibility of reversing the roles of ''data'' and ''message:'' with the scheme in Fig. 1, one could encode two qubits' worth of actual quantum data in the ''error syndrome,'' and a one-bit (or one-qubit) message in the ''logical qubit'' part of Eq. (4) (that is, the part involving α and β). What one gains with this scheme is efficiency, in the sense of a greater data-to-message information ratio. This can be made even greater by going to larger codes: for instance, any $[n,1]$ stabilizer code (encoding 1 logical qubit into n physical qubits) has $n-1$ generators, and measuring them determines the error syndrome; so the error syndrome can be extracted in $n-1$ qubits. With this ''reverse encoding'' approach, therefore, one could ''pack'' $n-1$ qubits worth of data in the error syndrome of every encoded logical qubit, and one qubit worth of ''message'' in the actual logical qubit.

As before, security would be provided by a secret key specifying on what basis each block is to be encoded. Since each block carries only one qubit of message, the key only has to be as long as the message itself; the ratio of length of key to length of data file (in bits, or qubits) is $1/(n-1)$, which could be made very small by using sufficiently large codes. If the total number of packages sent is t, the total number of possible different keys is 2^t, which suggests that the probability of successful tampering with or reading of the message is of the order of $1/2^t$. Clearly, though, this suggestion does not amount to a formal proof of security, and the scheme is only mentioned to give an indication of possibilities—in particular, for greater efficiency—beyond the ''straightforward'' scheme of the previous two subsections.

III. DISCUSSION AND CONCLUSIONS

The ideas proposed here provide some potentially useful ways to hide messages in quantum data for a variety of applications. In all cases, the data and the message are combined so that any change in one is likely to result in a change in the other. The method of Sec. II A is most useful when one is already considering using a QECC to protect the data against accidental errors, and the message to be sent (a sort of ''watermark'') is previously known to all the authorized parties; then the message can ''ride along'' in the error syndrome without compromising the efficacy of the code to correct errors. Note that, for data authentication purposes, the ''watermark'' should be quantum data, in order to provide protection against certain types of attacks.

As emphasized in the Introduction, the presentation here has been heuristic; I have not attempted to study the security features of the proposed schemes, for any specific application, in detail, nor to improve their efficiency beyond the suggestion, in Sec. II C, that smaller ratios of key to data file length might be possible. Nonetheless, as also mentioned in the Introduction, it is

interesting that the result exhibits a number of similarities to the more formal methodology for data authentication studied by Barnum *et al.*;[7] in particular, the use of QECCs (which under some circumstances could be used to provide additional protection for the data, as also remarked in Ref. 7) and the idea of hiding some secret ''key'' information in the error syndrome. It appears as if a careful study of the similarities and differences between the two schemes could be an enlightening and worthwhile task, but at present this must remain beyond the scope of the present manuscript.

In closing, to return, for a moment, to what constituted the original motivation for the present work namely, the idea of exploring the feasiblity of a sort of ''quantum steganography,'' these results do indicate how it may be possible, in principle, to combine a given set of quantum data and a quantum message in an inextricable way. It could be argued, however, that the final result does not look much like classical steganography, since we have ended up encrypting both the message and the data, whereas in classical steganography, instead, the ''data'' are, for the most part, openly visible (although subtly altered), and only the message is hidden. Yet, as I have mentioned above, it is not obvious to me how one could only ''slightly'' alter quantum information in order to accomplish this goal. In fact, it is one of the most interesting results proved in Ref. 7 that (unlike for classical data, where authentication and encryption are two different tasks) in order to authenticate quantum data one must, as it turns out, encrypt them ''almost perfectly.''

ACKNOWLEDGMENTS

I am grateful to Howard Barnum for very useful comments, and to the authors of Ref. 7 for sharing with me a preprint of their work. This research has been supported by the Army Research Office and the National Science Foundation.

[1] For an introduction and references, see, e.g., F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, in *Second Workshop on Information Hiding*, edited by D. Aucsmith, Vol. 1525 of *Lecture Notes in Computer Science*, p. 218; available electronically at http://www.cl.cam.ac.uk/Research/Security/.
[2] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf, Phys. Rev. Lett. **87**, 167902 (2001).
[3] D. Gottesman and I. L. Chuang, quant-ph/0105032.
[4] M. Curty and D. J. Santos, Phys. Rev. A **64**, 062309 (2001).
[5] M. Curty, D. J. Santos, E. Pérez, and P. García-Fernández, quant-ph/0108100 (to appear in Phys. Rev. A).
[6] D. W. Leung, quant-ph/0012077.
[7] H. Barnum, C. Crepeau, D. Gottesman, A. Smith, and A. Tapp, quant-ph/0205128.
[8] See, e.g., J. Preskill, Proc. R. Soc. London, Ser. A **454**, 385 (1998). For an introduction to quantum error-correcting codes, and *Quantum Computation and Quantum Information*, by M. A. Nielsen and I. L. Chuang (Cambridge University Press, Cambridge, 2000), Chap. 10, for the details.
[9] This particular ratio is obtained only if the message is classical (see the following subsection for how to embed a quantum message).