Simple Proof of Security of the BB84 Quantum Key Distribution Protocol

Peter W. Shor[1] and John Preskill[2]

[1]*AT&T Labs Research, Florham Park, New Jersey 07932*

[2]*Lauritsen Laboratory of High Energy Physics, California Institute of Technology, Pasadena, California 91125*

(Received 28 February 2000)

We prove that the 1984 protocol of Bennett and Brassard (BB84) for quantum key distribution is secure. We first give a key distribution protocol based on entanglement purification, which can be proven secure using methods from Lo and Chau's proof of security for a similar protocol. We then show that the security of this protocol implies the security of BB84. The entanglement purification based protocol uses Calderbank-Shor-Steane codes, and properties of these codes are used to remove the use of quantum computation from the Lo-Chau protocol.

PACS numbers: 03.67.Dd

Quantum cryptography differs from conventional cryptography in that the data are kept secret by the properties of quantum mechanics, rather than the conjectured difficulty of computing certain functions. The first quantum key distribution protocol, proposed in 1984 [1], is called BB84 after its inventors (Bennett and Brassard). In this protocol, the participants (Alice and Bob) wish to agree on a secret key about which no eavesdropper (Eve) can obtain significant information. Alice sends each bit of the secret key in one of a set of conjugate bases which Eve does not know, and this key is protected by the impossibility of measuring the state of a quantum system simultaneously in two conjugate bases. The original papers proposing quantum key distribution [1] proved it secure against certain attacks, including those feasible using current experimental techniques. However, for many years, it was not rigorously proven secure against an adversary able to perform any physical operation permitted by quantum mechanics.

Recently, three proofs of the security of quantum key distribution protocols have been discovered; however, none is entirely satisfactory. One proof [2], although easy to understand, has the drawback that the protocol requires a quantum computer. The other two [3,4] prove the security of a protocol based on BB84, and so are applicable to near-practical settings. However, both proofs are quite complicated. We give a simpler proof by relating the security of BB84 to entanglement purification protocols [5] and quantum error correcting codes [6]. This new proof also may illuminate some properties of previous proofs [3,4], and thus gives insight into them. For example, it elucidates why the rates obtainable from these proofs are related to rates for Calderbank-Shor-Steane (CSS) codes. The proof was in fact inspired by the observation that CSS codes are hidden in the inner workings of the proof given in [3].

We first review CSS codes and associated entanglement purification protocols. Quantum error correcting codes are subspaces of the Hilbert space $\mathbb{C}^{2^n}$ which are protected from errors in a small number of these qubits, so that any such error can be measured and subsequently corrected without disturbing the encoded state. A quantum CSS code Q on n qubits comes from two binary codes on n bits, C_1 and C_2, one contained in the other:

$$\{0\} \quad C_2 \quad C_1 \quad \mathbb{Z}_2^n$$

where $\mathbb{Z}_2^n$ is the binary vector space on n bits [6].

A set of basis states (which we call *code words*) for the CSS code subspace can be obtained from vectors $v \in C_1$ as follows:

$$|v\rangle \quad \frac{1}{|C_2|^{1/2}} \sum_{w \in C_2} |v + w\rangle . \tag{1}$$

If $v_1 - v_2 \in C_2$, then the code words corresponding to v_1 and v_2 are the same. Hence these code words correspond to cosets of C_2 in C_1, and this code protects a Hilbert space of dimension $2^{\dim C_1 - \dim C_2}$.

The above quantum code is equivalent to the dual code Q^* obtained from the two binary codes

$$\{0\} \quad C_1^\perp \quad C_2^\perp \quad \mathbb{Z}_2^n .$$

This equivalence can be demonstrated by applying the Hadamard transform

$$H \quad \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

to each encoding qubit. This transformation interchanges the bases $|0\rangle$, $|1\rangle$ and $|+\rangle$, $|-\rangle$, where $|+\rangle \quad \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $|-\rangle \quad \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$. It also interchanges the two subspaces corresponding to the codes Q and Q^*, although the code words [given by Eq. (1)] of Q and Q^* are not likewise interchanged.

We now make a brief technical detour to define some terms. The three Pauli matrices are

$$\sigma_x \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \sigma_y \quad \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

$$\sigma_z \quad \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} .$$

The matrix σ_x applies a bit flip error to a qubit, while σ_z applies a phase flip error. We denote the Pauli matrix σ_a acting on the kth bit of the CSS code by $\sigma_{a(k)}$ for

$a \in \{x\ y\ z\}$. For a binary vector s, we let

$$\sigma_a^{[s]} \quad \sigma_{a(1)}^{s_1} \otimes \sigma_{a(2)}^{s_2} \otimes \sigma_{a(3)}^{s_3} \otimes \cdots \otimes \sigma_{a(n)}^{s_n}$$

where σ_a^0 is the identity matrix and s_i is the ith bit of s. The matrices $\sigma_x^{[s]}$ ($\sigma_z^{[s]}$) have all eigenvalues ± 1.

In a classical error correcting code, correction proceeds by measuring the syndrome, which is done as follows. A *parity check* matrix H of a code is a basis of the dual vector space . Suppose that we transmit a code word v, which acquires errors to become $w \quad v + \epsilon$. The kth row r_k of the matrix H determines the kth bit of the syndrome for w, namely, $r_k \cdot w$ (mod2). The full syndrome is thus Hw. If the syndrome is 0, then $w \in$. Otherwise, the most likely value of the error ϵ can be calculated from the syndrome [7]. In our quantum CSS code, we need to correct both bit and phase errors. Let H_1 be a parity check matrix for the code $_1$, and H_2 one for the code $_2$. To calculate the syndrome for bit flips, we measure the eigenvalue of $\sigma_z^{[r]}$ for each row $r \in H_1$ (-1's and 1's of the eigenvalue correspond to 1's and 0's of the syndrome). To calculate the syndrome for phase flips, we measure the eigenvalue of $\sigma_x^{[r]}$ for each row $r \in H_2$. This lets us correct both bit and phase flips, and if we can correct up to t of each of these types of errors, we can also correct arbitrary errors on up to t qubits [6].

The useful property of CSS codes for demonstrating the security of BB84 is that the error correction for the phases is decoupled from that for the bit values, as shown above. General quantum stabilizer codes can similarly be turned into key distribution protocols, but these appear to require a quantum computer to implement.

If one requires that a CSS code corrects all errors on at most $t \quad \delta n$ qubits, the best codes that we know exist satisfy the quantum Gilbert-Varshamov bound. As the block length n goes to infinity, these codes asymptotically protect against δn bit errors and δn phase errors, and encode $[1 - 2H(2\delta)]n$ qubits, where H is the binary Shannon entropy $H(p) \quad -p \log_2(p) - (1-p)\log_2(1-p)$. In practice, it is better to require only that random errors are corrected with high probability. In this case, codes exist that correct δn random phase errors and δn random bit errors, and which encode $[1 - 2H(\delta)]n$ qubits.

We also need a description of the Bell basis. These are the four maximally entangled states:

$$\Psi^{\pm} \quad \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle) \qquad \Phi^{\pm} \quad \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle)$$

which form an orthogonal basis for the quantum state space of two qubits.

Finally, we introduce a class of quantum error correcting codes equivalent to Q, and parametrized by two n-bit binary vectors x and z. Suppose that Q is determined as above by $_1$ and $_2$. Then $Q_{x\,z}$ has basis vectors indexed by cosets of $_2$ in $_1$, and for $v \in _1$, the corresponding code word is

$$v \quad \frac{1}{|\ _2|^{1/2}} \sum_{w \in\ _2} (-1)^{z\cdot w} |x + v + w\rangle. \qquad (2)$$

Quantum error correcting codes and entanglement purification protocols are closely connected [5]; we now describe the entanglement purification protocol corresponding to the CSS code Q. For now, we assume that the codes $_1$ and $_2$ correct up to t errors and that Q encodes m qubits in n qubits. Suppose Alice and Bob share n pairs of qubits in a state close to $(\Phi^+)^{\otimes n}$. For the entanglement purification protocol, Alice and Bob separately measure the eigenvalues of $\sigma_z^{[r]}$ for each row $r \in H_1$ and $\sigma_x^{[r']}$ for each row $r' \in H_2$. Note that for these measurements to be performable simultaneously, they must all commute; $\sigma_z^{[r]}$ and $\sigma_x^{[r']}$ commute because the vector spaces $_1$ and $_2$ are orthogonal.

If Alice and Bob start with n perfect EPR pairs, measuring $\sigma_z^{[r]}$ for $r \in H_1$ and $\sigma_x^{[r']}$ for $r' \in H_2$ projects each of their states onto the code subspace $Q_{x\,z}$, where x and z are any binary vectors with $H_1 x$ and $H_2 z$ equal to the measured bit and phase syndromes, respectively. After projection, the state is $(\Phi^+)^{\otimes m}$ encoded by $Q_{x\,z}$.

Now, suppose that Alice and Bob start with a state close to $(\Phi^+)^{\otimes n}$. To be specific, suppose that all their EPR pairs are in the Bell basis, with t or fewer bit flips (Ψ^+ or Ψ^- pairs) and t or fewer phase flips (Φ^- or Ψ^- pairs). If Alice and Bob compare their measurements of $\sigma_z^{[r]}$ ($\sigma_x^{[r]}$), the rows r for which these measurements disagree give the bits which are 1 in the bit (phase) syndromes. From these syndromes, Alice and Bob can compute the locations of the bit and the phase flips, can correct these errors, and can then decode $Q_{x\,z}$ to obtain m perfect EPR pairs.

We will show that the following is a secure quantum key distribution protocol.

Protocol 1: Modified Lo-Chau.—(1) Alice creates $2n$ EPR pairs in the state $(\Phi^+)^{\otimes 2n}$. (2) Alice selects a random $2n$-bit string b, and performs a Hadamard transform on the second half of each EPR pair for which b is 1. (3) Alice sends the second half of each EPR pair to Bob. (4) Bob receives the qubits and publicly announces this fact. (5) Alice selects n of the $2n$ encoded EPR pairs to serve as check bits to test for Eve's interference. (6) Alice announces the bit string b, and which n EPR pairs are to be check bits. (7) Bob performs Hadamards on the qubits where b is 1. (8) Alice and Bob each measure their halves of the n check EPR pairs in the $|0\rangle, |1\rangle$ basis and share the results. If too many of these measurements disagree, they abort the protocol. (9) Alice and Bob make the measurements on their code qubits of $\sigma_z^{[r]}$ for each row $r \in H_1$ and $\sigma_x^{[r]}$ for each row $r \in H_2$. Alice and Bob share the results, compute the syndromes for bit and phase flips, and then transform their state so as to obtain m nearly perfect EPR pairs. (10) Alice and Bob measure the EPR pairs in the $|0\rangle, |1\rangle$ basis to obtain a shared secret key.

We now show that this protocol works. Namely, we show that the probability is exponentially small that Alice and Bob agree on a key about which Eve can obtain more than an exponentially small amount of information. We need a result of Lo and Chau [2] that if Alice and Bob share a state having fidelity $1 - 2^{-s}$ with $(\Phi^+)^{\otimes m}$, then Eve's mutual information with the key is at most $2^{-c} + 2^{O(-2s)}$ where $c \quad s - \log_2(2m + s + 1/\log_e 2)$.

For the proof, we use an argument based on one from Lo and Chau [2]. Let us calculate the probability that the test on the check bits succeeds while the entanglement purification on the code bits fails. We do this by considering the measurement that projects each of the EPR pairs onto the Bell basis.

We first consider the check bits. Note that for the EPR pairs where $b \quad 1$, Alice and Bob are effectively measuring them in the $|+\rangle, |-\rangle$ basis rather than the $|0\rangle, |1\rangle$ basis. Now, observe that

$$|\Psi^+\rangle\langle\Psi^+| + |\Psi^-\rangle\langle\Psi^-| \quad |01\rangle\langle 01| + |10\rangle\langle 10|$$
$$|\Phi^-\rangle\langle\Phi^-| + |\Psi^-\rangle\langle\Psi^-| \quad |+-\rangle\langle +-| + |-+\rangle\langle -+|.$$

These relations show that the rates of bit flip errors and of phase flip errors that Alice and Bob estimate from their measurements on check bits are the same as they would have estimated using the Bell basis measurement.

We next consider the measurements on the code bits. We want to show that the purification protocol applied to n pairs produces a state that is close to the encoded $(\Phi^+)^{\otimes m}$. The purification protocol succeeds perfectly acting on the space spanned by Bell pairs that differ from $(\Phi^+)^{\otimes n}$ by t or fewer bit flip errors and by t or fewer phase flips errors. Let Π denote the projection onto this space. Then if the protocol is applied to an initial density operator ρ of the n pairs, it can be shown that the final density operator ρ' approximates $(\Phi^+)^{\otimes m}$ with fidelity

$$F \quad \langle(\Phi^+)^{\otimes m}|\rho'|(\Phi^+)^{\otimes m}\rangle \quad \mathrm{tr}(\Pi\rho). \qquad (3)$$

Hence the fidelity is at least as large as the probability that t or fewer bit flip errors and t or fewer phase flip errors would have been found, if the Bell measurement had been performed on all n pairs.

Now, when Eve has access to the qubits, she does not yet know which qubits are check qubits and which are code qubits, so she cannot treat them differently. The check qubits that Alice and Bob measure thus behave like a classical random sample of the qubits. We are then able to use the measured error rates in a classical probability estimate; we find that probability of obtaining more than δn bit (phase) errors on the code bits and fewer than $(\delta - \epsilon)n$ errors on the check bits is asymptotically less than $\exp[-\frac{1}{4}\epsilon^2 n/(\delta - \delta^2)]$. We conclude that if Alice and Bob have greater than an exponentially small probability of passing the test, then the fidelity of Alice and Bob's state with $(\Phi^+)^{\otimes m}$ is exponentially close to 1.

We now show how to turn this Lo-Chau–type protocol into a quantum error correcting code protocol. Observe first that it does not matter whether Alice measures her check bits before or after she transmits half of each EPR pair to Bob, and similarly that it does not matter whether she measures the syndrome before or after this transmission. If she measures the check bits first, this is the same as choosing a random one of $|0\rangle, |1\rangle$. If she also measures the syndrome first, this is equivalent to transmitting m halves of EPR pairs encoded by the CSS code $Q_{x\,z}$ for two random vectors $x, z \in \ _2^n$. The vector x is determined by the syndrome measurements $\sigma_z^{[r]}$ for rows $r \in H_1$, and similarly for z. Alice can also measure her half of the encoded EPR pairs before or after transmission. If she measures them first, this is the same as choosing a random key k and encoding k using $Q_{x\,z}$. We thus obtain the following equivalent protocol.

Protocol 2: CSS Codes.—(1) Alice creates n random check bits, a random m-bit key k, and a random $2n$-bit string b. (2) Alice chooses n-bit strings x and z at random. (3) Alice encodes her key $|k\rangle$ using the CSS code $Q_{x\,z}$. (4) Alice chooses n positions (out of $2n$) and puts the check bits in these positions and the code bits in the remaining positions. (5) Alice applies a Hadamard transform to those qubits in the positions having 1 in b. (6) Alice sends the resulting state to Bob. Bob acknowledges receipt of the qubits. (7) Alice announces b, the positions of the check bits, the values of the check bits, and the x and z determining the code $Q_{x\,z}$. (8) Bob performs Hadamards on the qubits where b is 1. (9) Bob checks whether too many of the check bits have been corrupted, and aborts the protocol if so. (10) Bob decodes the key bits and uses them for the key.

Intuitively, the security of the protocol depends on the fact that for a sufficiently low error rate, a CSS code transmits the information encoded by it with very high fidelity, so that by the no-cloning principle very little information can leak to Eve.

We now give the final argument that turns the above protocol into BB84. First note that, since all Bob cares about are the bit values of the encoded key, and the string z is only used to correct the phase of the encoded qubits, Bob does not need z. This is why we use CSS codes: they decouple the phase correction from the bit correction. Let $k' \in \ _1$ be a binary vector that is mapped by Eq. (2) to the encoded key. Since Bob never uses z, we can assume that Alice does not send it. Averaging over z, we see that Alice effectively sends the mixed state

$$\frac{1}{2^n|\ _2|}\sum_z \left[\sum_{w_1}\sum_{w_2\in\ _2}(-1)^{(w_1+w_2)\cdot z}|k' + w_1 + x\rangle\langle k' + w_2 + x|\right] \quad \frac{1}{|\ _2|}\sum_{w\in\ _2}|k' + w + x\rangle\langle k' + w + x| \qquad (4)$$

which is equivalently the mixture of states $|k' + x + w\rangle$ with w chosen randomly in $_2$. Let us now look at the protocol as a whole. The error correction information Alice gives Bob is x, and Alice sends $|k' + x + w\rangle$ over the quantum channel. Over many iterations of the algorithm, these are random variables chosen uniformly in $_2^n$ with the constraint that their difference $k' + w$ is in $_1$. After Bob receives $k' + w + x + \epsilon$, he subtracts x, and corrects the result to a code word in $_1$, which is almost certain to be $k' + w$. The key is the coset of $k' + w$ over $_2$.

In the BB84 protocol given below, Alice sends $|v\rangle$ to Bob, with error correction information $u + v$. These are again two random variables uniform in $_2^n$, with the constraint that $u \in$ $_1$. Bob obtains $v + \epsilon$, subtracts $u + v$, and corrects the result to a code word in $_1$, which with high probability is u. The key is then the coset $u +$ $_2$. Thus, the two protocols are completely equivalent.

Protocol 3: BB84.— (1) Alice creates $(4 + \delta)n$ random bits. (2) Alice chooses a random $(4 + \delta)n$-bit string b. For each bit, she creates a state in the $|0\rangle$, $|1\rangle$ basis (if the corresponding bit of b is 0) or the $|+\rangle$, $|-\rangle$ basis (if the bit of b is 1). (3) Alice sends the resulting qubits to Bob. (4) Bob receives the $(4 + \delta)n$ qubits, measuring each in the $|0\rangle$,$|1\rangle$ or the $|+\rangle$,$|-\rangle$ basis at random. (5) Alice announces b. (6) Bob discards any results where he measured a different basis than Alice prepared. With high probability, there are at least $2n$ bits left (if not, they abort the protocol). Alice decides randomly on a set of $2n$ bits to use for the protocol, and chooses at random n of these to be check bits. (7) Alice and Bob announce the values of their check bits. If too few of these values agree, they abort the protocol. (8) Alice announces $u + v$, where v is the string consisting of the remaining noncheck bits, and u is a random code word in $_1$. (9) Bob subtracts $u + v$ from his code qubits, $v + \epsilon$, and corrects the result, $u + \epsilon$, to a code word in $_1$. (10) Alice and Bob use the coset of $u +$ $_2$ as the key.

There are a few loose ends that need to be tied up. The protocol given above uses binary codes $_1$ and $_2$ with large minimum distance, and thus can obtain rates given by the quantum Gilbert-Varshamov bound for CSS codes [6]. To reach the better Shannon bound for CSS codes, we need to use codes for which a random small set of phase errors and bit errors can almost always be corrected. To prove that the protocol works in this case, we need to ensure that the errors are indeed random. We do this by adding a step where Alice scrambles the qubits using a random permutation π before sending them to Bob, and a step after Bob acknowledges receiving the qubits where Alice sends π to Bob and he unscrambles the qubits. This can work as long as the measured bit and phase error rates are less than 11%, the point at which the Shannon rate $1 - 2H(\delta)$ hits 0.

For a practical key distribution protocol we need the classical code $_1$ to be efficiently decodeable. As is shown in [3], we can let $_2$ be a random subcode of an efficiently decodeable code $_1$, and with high probability obtain a good code $_2$. While known efficiently decodeable codes do not meet the Shannon bound, they come fairly close.

A weakness in both the proof given in this paper and the proofs in [3,4] is that they do not apply if Alice sometimes inadvertently sends two or more identical copies of her qubit instead of just one copy. A proof avoiding this difficulty was recently discovered by Ben-Or [8]; it shows that any source sufficiently close to a single-photon source is still secure. However, most experimental quantum key distribution systems use weak coherent sources, and no currently known proof covers this case.

The authors thank Michael Ben-Or, Eli Biham, Hoi-Kwong Lo, Dominic Mayers, and Tal Mor for explanations of and informative discussions about their security proofs. We also thank Ike Chuang, Daniel Gottesman, Alexei Kitaev, and Michael Nielsen for their discussions and suggestions, which greatly improved this paper. Part of this research was done while P. W. S. was visiting Caltech. This work has been supported in part by the Department of Energy under Grant No. DE-FG03-92-ER40701, and by DARPA through Caltech's Quantum Information and Computation (QUIC) project administered by the Army Research Office.

[1] C. H. Bennett and G. Brassard, in *Proceedings of the IEEE International Conference on Computers, Systems and Signal Processing, Bangalore, India, 1984* (IEEE, New York, 1984), pp. 175–179; IBM Tech. Discl. Bull. **28**, 3153–3163 (1985).

[2] H.-K. Lo and H. F. Chau, Science **283**, 2050–2056 (1999).

[3] D. Mayers, J. Assoc. Comput. Mach. (to be published), quant-ph/9802025; preliminary version in *Advances in Cryptology — Proceedings of Crypto '96* (Springer-Verlag, New York, 1996), pp. 343–357.

[4] E. Biham, M. Boyer, P. O. Boykin, T. Mor, and V. Roychowdhury, in *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing* (ACM Press, New York, 2000), pp. 715–724.

[5] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, Phys. Rev. A **54**, 3824–3851 (1996).

[6] A. R. Calderbank and P. Shor, Phys. Rev. A **54**, 1098–1105 (1996); A. M. Steane, Proc. R. Soc. London A **452**, 2551–2577 (1996).

[7] This calculation may be quite difficult, but for now we ignore this practical complication.

[8] M. Ben-Or and I. Bregman (unpublished).