Quantum Fingerprinting

Harry Buhrman,[1,*] Richard Cleve,[2,†] John Watrous,[2,‡] and Ronald de Wolf[1,§]

[1]*CWI, P.O. Box 94709, Amsterdam, The Netherlands*
and University of Amsterdam, Amsterdam, The Netherlands
[2]*Department of Computer Science, University of Calgary, Calgary, Alberta, Canada T2N 1N4*

(Received 19 April 2001; published 26 September 2001)

Classical fingerprinting associates with each string a shorter string (its *fingerprint*), such that any two distinct strings can be distinguished with small error by comparing their fingerprints alone. The fingerprints cannot be made exponentially smaller than the original strings unless the parties preparing the fingerprints have access to correlated random sources. We show that fingerprints consisting of *quantum* information *can* be made exponentially smaller than the original strings without any correlations or entanglement between the parties. This implies an exponential quantum/classical gap for the equality problem in the simultaneous message passing model of communication complexity.

DOI: 10.1103/PhysRevLett.87.167902 PACS numbers: 03.67.Hk, 03.65.Ta, 03.67.Lx

Fingerprinting can be a useful mechanism for determining if two strings are the same: each string is associated with a much shorter fingerprint and comparisons between strings are made in terms of their fingerprints alone. This can lead to savings in the communication and storage of information.

The notion of fingerprinting arises naturally in the setting of *communication complexity* (see [1] for a survey). The particular model of communication complexity that we consider in this Letter is called the *simultaneous message passing* model, which was introduced by Yao [2] in his original paper on communication complexity. In this model, two parties — Alice and Bob — receive inputs x and y, respectively, and are not permitted to communicate with one another directly. Rather they each send a message to a third party, called the *referee*, who determines the output of the protocol based solely on the messages sent by Alice and Bob. The collective goal of the three parties is to cause the protocol to output the correct value of some function $f(x,y)$ while minimizing the amount of communication from Alice and Bob to the referee. For the *equality* problem, the function is

$$f(x,y) = \begin{cases} 1, & \text{if } x = y\,, \\ 0, & \text{if } x \neq y\,. \end{cases} \tag{1}$$

The problem can, of course, be trivially solved if Alice sends x and Bob sends y to the referee, who can then compute $f(x,y)$. However, the cost of this protocol is high; if x and y are n-bit strings, then a total of $2n$ bits are communicated. If Alice and Bob instead send *fingerprints* of x and y, which may each be considerably shorter than x and y, the cost can be reduced significantly. The question we are interested in is how much the size of the fingerprints can be reduced.

If Alice and Bob share a random $O(\log_2(n))$-bit key, then the fingerprints need only be of constant length if we allow a small probability of error; a brief sketch of this follows. A binary error-correcting code is used, which can be represented as a function $E\colon \{0,1\}^n \rightarrow \{0,1\}^m$, where $E(x)$ is the code word associated with $x \in \{0,1\}^n$. There exist error-correcting codes (Justesen codes, for instance) with $m = cn$ such that the Hamming distance between any two distinct code words $E(x)$ and $E(y)$ (i.e., the number of bit positions where they differ) is at least $(1-\delta)m$, where c and δ are positive constants. For the particular case of Justesen codes, we may choose any $c > 2$ and we will have $\delta < 9/10 + 1/(15c)$ (for sufficiently large n) [3]. Now, for $x \in \{0,1\}^n$ and $i \in \{1,2,\ldots,m\}$, let $E_i(x)$ denote the ith bit of $E(x)$. The shared key is a random $i \in \{1,2,\ldots,m\}$ [consisting of $\log_2(n) + O(1)$ bits]. Alice and Bob, respectively, send the bits $E_i(x)$ and $E_i(y)$ to the referee, who then outputs 1 if and only if $E_i(x) = E_i(y)$. If $x = y$, then $E_i(x) = E_i(y)$, so then the outcome is correct. If $x \neq y$, then the probability that $E_i(x) = E_i(y)$ is at most δ, so the outcome is correct with probability $1-\delta$. The error probability can be reduced from δ to any $\varepsilon > 0$ by having Alice and Bob send $O(\log_2(1/\varepsilon))$ independent random bits of the code words $E(x)$ and $E(y)$ to the referee. In this case, the length of each fingerprint is $O(\log_2(1/\varepsilon))$ bits.

One disadvantage of the above scheme is that is requires overhead in creating and maintaining a shared key. Moreover, once the key is distributed, it may be necessary to store it securely until the inputs are obtained. This is because, for every fixed key value, there are distinct inputs x and y on which the protocol gives the incorrect output 1. Therefore, an adversary who uses the shared key as prior information can perform the task of fooling the protocol into incorrectly outputting the value 1.

Yao (Ref. [2] Section 4.D) posed as an open problem the question of what happens in this model if Alice and Bob do not have a shared key. Ambainis [4] proved that fingerprints of $O(\sqrt{n})$ bits suffice if we allow a small error probability (see also [5–7]). Note that in this setting Alice and Bob still have access to random bits, but there are no correlations between each others' random bits. Subsequently, Newman and Szegedy [7] proved the above is optimal in that the length of the fingerprints must scale at least proportionally to $\sqrt{n}$. Babai and Kimmel [5] later showed that probabilistic and deterministic communication complexity can be at most quadratically far apart for *any* function in the simultaneous message passing model, which also implies the $\sqrt{n}$ lower bound. Babai and Kimmel attribute a simplified proof of this fact to Jean Bourgain and Avi Wigderson.

We consider the problem where Alice and Bob's fingerprints can consist of *quantum* information. Alice and Bob are still restricted to have no shared key (or entanglement) between them. We show that $O(\log_2(n))$-qubit fingerprints are sufficient to solve the equality problem in this setting—an exponential improvement over the $\sqrt{n}$-bound for the comparable classical case. Our method is to set the 2^n fingerprints to quantum states whose pairwise inner products are bounded below 1 in absolute value and to use a measurement that identifies identical fingerprints and distinguishes distinct fingerprints with good probability. This gives a simultaneous message passing protocol for equality in the obvious way: Alice and Bob send the fingerprints of their respective inputs to the referee, who then performs the measurement that checks if the fingerprints are equal or distinct.

The fact that quantum systems contain large sets of nearly orthogonal states—sets of 2^n states that are nearly orthogonal pairwise in $O(\log_2(n))$-qubit systems—is well known. For example, it is noted in [8], where it is shown that these nearly orthogonal sets of states cannot be utilized to solve certain coding problems much more efficiently than possible with classical information. Our results are perhaps the first demonstration that nearly orthogonal sets of quantum states can be used to perform a natural information processing task significantly more efficiently than possible with classical information.

To explicitly construct a large set of nearly orthogonal quantum states, assume that for fixed $c > 1$ and $0 < \delta < 1$ we have an error correcting code E: $\{0,1\}^n \rightarrow \{0,1\}^m$ for each n, where $m = cn$ and such that the distance between distinct code words $E(x)$ and $E(y)$ is at least $(1-\delta)m$. For instance, we may use the codes discussed previously in the classical shared-key protocol. Now, for each $x \in \{0,1\}^n$, define the $(\log_2(m)+1)$-qubit state

$$|h_x\rangle = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} |i\rangle |E_i(x)\rangle . \tag{2}$$

Since two distinct code words can be equal in at most δm positions, for any $x \neq y$ we have $\langle h_x | h_y \rangle \leq \delta m/m = \delta$. Thus we have 2^n different $(\log_2(n) + O(1))$-qubit states, and each pair of them has an inner product with an absolute value at most δ.

The simultaneous message passing protocol for the equality problem works as follows. When given n-bit inputs x and y, respectively, Alice and Bob send fingerprints $|h_x\rangle$ and $|h_y\rangle$ to the referee. Then the referee must distinguish between the case where the two states received (call them $|\phi\rangle$ and $|\psi\rangle$) are identical or have an inner product at most δ in absolute value. This is accomplished with one-sided error probability by the procedure that measures and outputs the first qubit of the state

$$(H \otimes I)(\text{c-SWAP})(H \otimes I)|0\rangle|\phi\rangle|\psi\rangle . \tag{3}$$

Here H is the Hadamard transform, which maps $|b\rangle \rightarrow \frac{1}{\sqrt{2}}(|0\rangle + (-1)^b|1\rangle)$, SWAP is the operation $|\phi\rangle|\psi\rangle \rightarrow |\psi\rangle|\phi\rangle$, and c-SWAP is the controlled-SWAP (controlled by the first qubit). Figure 1 illustrates this. Tracing through the execution of this circuit, the final state before the measurement is

$$\tfrac{1}{2}|0\rangle(|\phi\rangle|\psi\rangle + |\psi\rangle|\phi\rangle) + \tfrac{1}{2}|1\rangle(|\phi\rangle|\psi\rangle - |\psi\rangle|\phi\rangle) . \tag{4}$$

Measuring the first qubit of this state produces outcome 1 with probability $(1 - |\langle\phi|\psi\rangle|^2)/2$. This probability is 0 if $x = y$ and is at least $(1-\delta^2)/2 > 0$ if $x \neq y$. Thus, the test determines which case holds with one-sided error probability $(1+\delta^2)/2$.

The error probability of the test can be reduced to any $\varepsilon > 0$ by setting the fingerprint of $x \in \{0,1\}^n$ to $|h_x\rangle^{\otimes k}$ for a suitable $k \in O(\log_2(1/\varepsilon))$. From such fingerprints, the referee can independently perform the test in Fig. 1 k times, resulting in an error probability below ε. In this case, the length of each fingerprint is $O(\log_2(n)\log_2(1/\varepsilon))$. In summary, we have shown the following.

Theorem 1.—There exists a quantum simultaneous message passing protocol for the equality problem with small error probability and $O(\log_2(n))$ qubits of communication [contrasting with $\Theta(\sqrt{n})$ bits classically].

It is worth considering what goes wrong if one tries to simulate the above quantum protocol using classical mixtures in place of quantum superpositions. In such a protocol, Alice and Bob send $(i, E_i(x))$ and $(j, E_j(y))$, respectively, to the referee for *independent* random uniformly distributed $i, j \in \{1, 2, \ldots, m\}$. If it should happen that $i = j$, then the referee can make a statistical inference about whether or not $x = y$. But $i = j$ occurs with probability only $O(1/n)$, and in the case where $i \neq j$, the referee will not be able to determine whether $x = y$ with good probability, as shown by the $\sqrt{n}$ lower bound of [7]. The distinguishing test in Fig. 1 can be viewed as a quantum operation that has no analogous classical probabilistic counterpart.

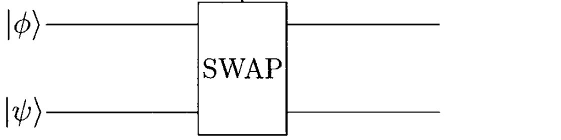

FIG. 1. Quantum circuit to test if $|\phi\rangle = |\psi\rangle$ or $|\langle\phi|\psi\rangle| \leq \delta$.

Our quantum protocol for equality in the simultaneous message model uses $O(\log_2(n))$-qubit fingerprints for any constant error probability. Is it possible to use fewer qubits? In fact, without a shared key, logarithmic-length fingerprints are necessary. This is because any k-qubit quantum state can be specified within exponential precision with $O(k2^k)$ classical bits. Therefore the existence of a k-qubit quantum protocol implies the existence of an $O(k2^k)$-bit (deterministic) classical protocol. From this we can infer that $k \geq \log_2(n) - O[\log_2(\log_2(n))]$.

We next consider some efficiency improvements to our fingerprinting scheme. It can be shown that the aforementioned method uses $k(\log_2(n) + O(1))$ qubit fingerprints to attain an error probability slightly more than $(9/10)^k$. First we note that the construction of nearly orthogonal states can be improved by using a better error-correcting code. Using a probabilistic argument (see, e.g., [9]), it can be shown that, for an arbitrarily small $\delta > 0$, there exists an error-correcting code E: $\{0,1\}^n \rightarrow \{0,1\}^m$ with $m \leq n/\delta^c$ (for some constant c) such that the Hamming distance between any two distinct code words $E(x)$ and $E(y)$ is between $(1 - \delta)m/2$ and $(1 + \delta)m/2$. If a set S of 2^n m-bit strings is chosen at random, then the probability that there is a pair of strings in S whose Hamming distance deviates from $m/2$ by more than δm is less than 1. This shows that there exists a set S with the right properties. Note that this existence proof does not yield an explicit construction of the code; however, Guruswami and Smith [10] recently pointed out to us that explicit constructions of such codes can be obtained from results in [11,12]. Given such a code, the $\log_2(m)$-qubit fingerprint of $x \in \{0,1\}^n$ can be set to

$$|h_x\rangle = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (-1)^{E_i(x)} |i\rangle \tag{5}$$

to yield the following theorem.

Theorem 2.—For every n and $\delta > 0$ one can construct a set $\{|h_x\rangle$: $x \in \{0,1\}^n\}$ of states of $\log_2(n) + O(\log_2(1/\delta))$ qubits, such that $|\langle h_x|h_y\rangle| \leq \delta$ whenever $x \neq y$.

The above construction yields fingerprints that are arbitrarily close to orthogonal—their pairwise inner products are within any $\delta > 0$ of 0. This results in a distinguishing measurement (Fig. 1) that errs with probability $(1 + \delta^2)/2$—slightly more than 1/2. To reduce the error probability to an arbitrarily small $\varepsilon > 0$, recall that the method we proposed is to construct k copies of each fingerprint, which can then be measured in pairs independently. The result is an error probability of $((1 + \delta^2)/2)^k$, which is approximately $1/2^k$ when δ is small. We now show that an alternate measurement results in an error probability close to $\sqrt{\pi k}\,((1 + \delta)/2)^{2k}$, which is approximately $\sqrt{\pi k}/4^k$ when δ is small. This is a near-quadratic reduction in the error probability resulting from a k-copy fingerprint consisting of $k(\log_2(n) + O(1))$ qubits.

The improved measurement works as follows. Let $R_1, \ldots, R_{2k}$ be registers that initially contain $|\phi\rangle, \ldots, |\phi\rangle$, $|\psi\rangle, \ldots, |\psi\rangle$ (k copies of each). Let $s = (2k)!$ and $\sigma_0, \sigma_1, \ldots, \sigma_{s-1}$ be an enumeration of all the permutations on $2k$ items, where σ_0 is the identity permutation. Let P be an s-dimensional register initialized to $|0\rangle$. Let F be any transformation satisfying

$$F\colon |0\rangle \mapsto \frac{1}{\sqrt{s}} \sum_{i=0}^{s-1} |i\rangle, \tag{6}$$

such as the s-dimensional quantum Fourier transform. Since s is a smooth number [i.e., its prime factors are all $O(\log_2(s))$], the construction in [13] implies that F can be computed exactly with a polynomial number of basic operations. The distinguishing procedure is as follows: (1) Apply F to register P. (2) Apply permutation σ_i to registers $R_1, \ldots, R_{2k}$, conditioned on the value of P being $|i\rangle$. (3) Apply $F^\dagger$ to P and measure the final state. If P contains 0, then answer *equal,* otherwise *not equal.* This procedure corresponds to a projection onto the *symmetric subspace* for registers $R_1, \ldots, R_{2k}$, as explained in [14]. The state after step 2 is

$$\frac{1}{\sqrt{s}} \sum_{i=0}^{s-1} |i\rangle \sigma_i(|\phi\rangle \cdots |\phi\rangle |\psi\rangle \cdots |\psi\rangle), \tag{7}$$

where $\sigma_i(|\phi\rangle \cdots |\phi\rangle |\psi\rangle \cdots |\psi\rangle)$ means we permute the contents of the $2k$ registers according to σ_i.

Case 1: $|\phi\rangle = |\psi\rangle$. In this case the permutation of the registers does absolutely nothing, so the procedure answers *equal* with certainty.

Case 2: $|\langle\phi|\psi\rangle| < \delta$. The probability of answering *equal* is the squared norm of the vector obtained by applying the projection $|0\rangle\langle 0| \otimes I$ to the final state:

$$\left\| \frac{1}{\sqrt{s}} \sum_{i=0}^{s-1} \langle 0|F^\dagger|i\rangle \sigma_i(|\phi\rangle \cdots |\phi\rangle |\psi\rangle \cdots |\psi\rangle) \right\|^2 \tag{8}$$

$$= \left\| \frac{1}{s} \sum_{i=0}^{s-1} \sigma_i(|\phi\rangle \cdots |\phi\rangle |\psi\rangle \cdots |\psi\rangle) \right\|^2 \tag{9}$$

$$= \frac{(k!)^2}{(2k)!} \sum_{j=0}^{k} \binom{k}{j}^2 \delta^{2j} \tag{10}$$

$$\leq \frac{(k!)^2}{(2k)!} (1 + \delta)^{2k} \sim \sqrt{\pi k} \left(\frac{1 + \delta}{2}\right)^{2k}. \tag{11}$$

In summary, we have shown the following.

Theorem 3.—The above procedure, on input $|\phi\rangle^{\otimes k}$ and $|\psi\rangle^{\otimes k}$ such that either $|\phi\rangle = |\psi\rangle$ or $|\langle\phi|\psi\rangle| \leq \delta$, decides which of the two is the case with error $O(\sqrt{k}\,(\frac{1+\delta}{2})^{2k})$.

The above procedure can be viewed as a solution to a more general *state distinguishing* problem defined as follows. The input is k copies of each of two quantum states $|\phi\rangle$ and $|\psi\rangle$ that are arbitrary subject to the condition that the two states are either identical or have inner product bounded in absolute value by some given $\delta < 1$.

The goal is to distinguish between the two cases with as high probability as possible. The above procedure solves the state distinguishing problem with error probability $\sqrt{\pi k}\,[(1 + \delta)/2]^{2k}$, and it can be shown that, in general, the error probability cannot be less than $(1/4)$ $[(1 + \delta)/2]^{2k}$. The idea behind this lower bound is to consider the pairs of states $|\phi_1\rangle = |\psi_1\rangle = |0\rangle$ and $|\phi_2\rangle = \cos(\theta/2)\,|0\rangle + \sin(\theta/2)\,|1\rangle$ and $|\psi_2\rangle = \cos(\theta/2)\,|0\rangle - \sin(\theta/2)\,|1\rangle$, where $\theta = \cos^{-1}(\delta)$. Clearly, $|\phi_1\rangle = |\psi_1\rangle$ and $\langle\phi_2|\psi_2\rangle = \delta$. A state distinguishing procedure must distinguish between $|a\rangle = |\phi_1\rangle^{\otimes k} \otimes |\psi_1\rangle^{\otimes k}$ and $|b\rangle = |\phi_2\rangle^{\otimes k} \otimes |\psi_2\rangle^{\otimes k}$. Since $\langle\phi_1|\phi_2\rangle = \langle\psi_1|\psi_2\rangle = \cos(\theta/2)$, it follows that $\langle a|b\rangle = \cos^{2k}(\theta/2) = [(1 + \cos\theta)/2]^k = [(1 + \delta)/2]^k$. It is known that the optimal procedure distinguishing between two states with inner product $\cos\alpha$ has error probability $(1 - \sin\alpha)/2 \geq (1/4)\cos^2\alpha$ [15]. Therefore any state distinguisher has error probability at least $(1/4)[(1 + \delta)/2]^{2k}$. Note that this lower bound for state distinguishing concerns a problem that is more general than the problem of distinguishing between *fingerprints,* because, in the case of fingerprints, the states are from a known set of only 2^n possibilities.

Finally, returning to the fingerprinting scenario, we consider the case where Alice and Bob have a shared *quantum* key, consisting of $O(\log_2(n))$ Bell states, but are required to output *classical* strings as fingerprints. Is there any sense in which a quantum key can result in improved performance over the case of a classical key? We observe that results in [16] imply an improvement in the particular setting where the fingerprinting scheme must be exact (i.e., the error probability is 0) and where there is a restriction on the inputs that either $x = y$ or the Hamming distance between x and y is $n/2$. Under this restriction, any classical scheme with a shared key would still require fingerprints of length linear in n. On the other hand, there is a scheme with a shared quantum key of $O(\log_2(n))$ Bell states that requires fingerprints of length only $O(\log_2(n))$ bits. See [16] (whose results are partly based on results in [17,18]) for details. It should be noted that if the exactness condition is relaxed to one where the error probability must be $O(1/n^c)$ (for a constant c) then there also exists a classical scheme with classical keys and fingerprints of length $O(\log_2(n))$.

We thank Andris Ambainis, Venkatesan Guruswami, Ashwin Nayak, John Preskill, and Adam Smith for helpful comments about this work. Some of this research took place while R. C. was at the CWI and while H. B. and R. C. were at Caltech, and the hospitality of these institutions is gratefully acknowledged. H. B. and R. d. W. are partially supported by the EU fifth framework project QAIP, IST-1999-11234. R. C. and J. W. are partially supported by Canada's NSERC.

*Email address: buhrman@cwi.nl
†Email address: cleve@cpsc.ucalgary.ca
‡Email address: jwatrous@cpsc.ucalgary.ca
§Email address: rdewolf@cwi.nl

[1] E. Kushilevitz and N. Nisan, *Communication Complexity* (Cambridge University Press, Cambridge, 1997).

[2] A. C.-C. Yao, in *Proceedings of the 11th Annual ACM Symposium on Theory of Computing* (ACM, New York, 1979), p. 209.

[3] J. Justesen, IEEE Trans. Inf. Theory **18**, 652 (1972).

[4] A. Ambainis, Algorithmica **16**, 298 (1996).

[5] L. Babai and P. G. Kimmel, in *Proceedings of the 12th Annual IEEE Conference on Computational Complexity* (IEEE, Los Alamitos, California, 1997), p. 239.

[6] I. Kremer, N. Nisan, and D. Ron, in *Proceedings of the 27th Annual ACM Symposium on Theory of Computing* (ACM, New York, 1995), p. 596.

[7] I. Newman and M. Szegedy, in *Proceedings of the 28th Annual ACM Symposium on Theory of Computing* (ACM, New York, 1996), p. 561.

[8] A. Ambainis, A. Nayak, A. Ta-Shma, and U. Vazirani, in *Proceedings of the 31st Annual ACM Symposium of Theory of Computing* (ACM, New York, 1999), p. 376.

[9] N. Alon and J.H. Spencer, *The Probabilistic Method* (Wiley-Interscience, New York, 1992).

[10] V. Guruswami and A. Smith (private communication).

[11] N. Alon, J. Bruck, J. Naor, M. Naor, and R. Roth, IEEE Trans. Inf. Theory **38**, 509 (1992).

[12] J. Bierbrauer and H. Schellwat, in *Advances in Cryptology—CRYPTO 2000 Proceedings* (Springer-Verlag, New York, 2000), p. 533.

[13] R. Cleve, "A note on computing Fourier transforms by quantum programs" (unpublished).

[14] A. Barenco, A. Berthiaume, D. Deutsch, A. Ekert, R. Jozsa, and C. Macchiavello, SIAM J. Comput. **26**, 1541 (1997).

[15] C. W. Helstrom, Information and Control **10**, 254 (1967).

[16] G. Brassard, R. Cleve, and A. Tapp, Phys. Rev. Lett. **83**, 1874 (1999).

[17] H. Buhrman, R. Cleve, and A. Wigderson, in *Proceedings of the 30th Annual ACM Symposium on Theory of Computing* (ACM, New York, 1998), p. 63.

[18] P. Frankl and V. Rödl, Trans. Am. Math. Soc. **300**, 259 (1987).