Quantum fingerprinting

Harry Buhrman[1], Richard Cleve[2], John Watrous[2], Ronald de Wolf[1]
[1] *CWI, P.O. Box 94709, Amsterdam, The Netherlands* {buhrman,rdewolf}@cwi.nl
[2] *Department of Computer Science, University of Calgary, Calgary, Alberta, Canada T2N 1N4*
{cleve,jwatrous}@cpsc.ucalgary.ca

Classical fingerprinting associates with each string a shorter string (its *fingerprint*), such that, with high probability, any two distinct strings can be distinguished by comparing their fingerprints alone. The fingerprints can be exponentially smaller than the original strings if the parties preparing the fingerprints share a random key, but not if they only have access to uncorrelated random sources. In this paper 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. In our scheme, the fingerprints are exponentially shorter than the original strings and a measurement distinguishes between the fingerprints of any two distinct strings. Our scheme implies an exponential quantum/classical gap for the equality problem in the simultaneous message passing model of communication complexity. We optimize several aspects of our scheme.

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 [11]). The particular model of communication complexity that we consider in this paper is called the *simultaneous message passing* model, which was introduced by Yao [16] 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 information that Alice and Bob send to the referee.

For the *equality* problem, the function is simply

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

The problem can of course be trivially solved if Alice sends x and Bob sends y to the referee, who can then simply 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 $O(\log n)$-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 $E : \{0,1\}^n \to \{0,1\}^m$, where $E(x)$ is the codeword 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 codewords $E(x)$ and $E(y)$ 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 $\delta < 9/10 + 1/(15c)$ (assuming n is sufficiently large). For further information on Justesen codes, see Justesen [10] and MacWilliams and Sloane [13, Chapter 10]. Now, for $x \in \{0,1\}^n$ and $i \in \{1,2,\ldots,m\}$, let $E_i(x)$ denote the i^{th} bit of $E(x)$. The shared key is a random $i \in \{1,2,\ldots,m\}$ (which consists 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 $O(\log(1/\varepsilon))$ independent random bits of the codewords $E(x)$ and $E(y)$ to the referee. In this case, the length of each fingerprint is $O(\log(1/\varepsilon))$ bits.

One disadvantage of the above scheme is that it requires overhead in creating and maintaining a shared key. Moreover, once the key is distributed, it must be stored securely until the inputs are obtained. This is because an adversary who knows the value of the key can easily choose inputs x and y such that $x \neq y$ but for which the output of the protocol always indicates that $x = y$.

Yao [16, 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 [1] proved that fingerprints of $O(\sqrt{n})$ bits suffice if we allow a small error probability (see also [6,12,15]). 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 [15] proved the above is optimal in that the length of the fingerprints must scale at least proportionally to $\sqrt{n}$. Babai and Kimmel [6] 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 $O(\log n)$-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 special 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 $O(\log n)$-qubit systems—is well known. For example, it is noted in [2], 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 $E : \{0,1\}^n \to \{0,1\}^m$ for each n, where $m = cn$ and such that the distance between distinct codewords $E(x)$ 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 $(\log(m)+1)$-qubit state $|h_x\rangle$ as

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

Since two distinct codewords 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 inner-product with 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 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 $|b\rangle \to \frac{1}{\sqrt{2}}(|0\rangle + (-1)^b|1\rangle)$, SWAP is the operation $|\phi\rangle|\psi\rangle \to |\psi\rangle|\phi\rangle$ and c-SWAP is the controlled-SWAP (controlled by the first qubit). A quantum circuit for this procedure is illustrated in Figure 1.

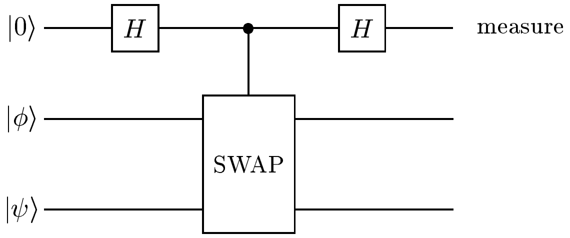

FIG. 1. Circuit to test if $|\phi\rangle = |\psi\rangle$ or $|\langle\phi|\psi\rangle| \leq \delta$

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 $\frac{1}{2} - \frac{1}{2}|\langle\phi|\psi\rangle|^2$. 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 $k \in O(\log(1/\varepsilon))$. From such fingerprints, the referee can independently perform the test in Figure 1 k times, resulting in an error probability below ε. In this case, the length of each fingerprint is $O((\log n)(\log(1/\varepsilon))$.

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 the ability of the referee to make an inference when $i \neq j$ seems difficult. For many error-correcting codes, no inference whatsoever about $x = y$ is possible when $i \neq j$ and the lower bound in [15] implies that no error-correcting code enables inferences to be made when $i \neq j$ with error probability bounded below 1. The distinguishing test in Figure 1 can be viewed as a quantum operation that has no analogous classical probabilistic counterpart.

Our quantum protocol for equality in the simultaneous message model uses $O(\log n)$-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 $k \geq c \log_2 n$ for some constant $c > 0$.

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, it can be shown that, for an arbitrarily small $\delta > 0$, there exists an error-correcting code $E : \{0,1\}^n \to \{0,1\}^m$ with $m \in \log_2(n) + O(\log(1/\delta))$ such that the Hamming distance between any two distinct codewords $E(x)$ and $E(y)$ is between $\frac{1}{2}(1-\delta)m$ and $\frac{1}{2}(1+\delta)m$. The idea is to show that if 2^n m-bit strings are chosen randomly then the probability that any two of them has Hamming distance more than $\frac{1}{2}\delta m$ away from $\frac{1}{2}m$ is less than 1. An extensive survey of such probabilistic arguments can be found in [3]. Note that this existence proof does not yield an explicit construction of the code; however, given such a code, the $\log m$-qubit fingerprint of $x \in \{0,1\}^n$ can be defined as

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

It is straightforward to show that, for any $x \neq y$, $|\langle h_x | h_y \rangle| \leq \delta$.

The above construction yields fingerprints that are arbitrarily close to orthogonal—their pairwise inner-products are within any $\delta > 0$ of 0 provided the qubit-length of the fingerprints is set to $\log_2(n) + O(1/\delta)$. This results in a distinguishing measurement (Figure 1) that errs with probability of $(1+\delta^2)/2$—slightly more than $\frac{1}{2}$. 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 for the state distinguishing problem 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 P be a resister whose classical states include encodings of all the permutations in S_{2k}. Let 0 encode the identity permutation and let P be initialized to 0. Let F be any transformation satisfying

$$F : |0\rangle \mapsto \frac{1}{\sqrt{(2k)!}} \sum_{\sigma \in S_{2k}} |\sigma\rangle. \tag{6}$$

Such a transformation can easily be computed in polynomial time.

The distinguishing procedure operates as follows:

1. Apply F to register P.
2. Apply a conditional permutation on the contents of registers $R_1, \ldots, R_{2k}$, conditioned on the permutation specified in P.
3. Apply $F^\dagger$ to P and measure the final state. If P contains 0 then answer *equal*, otherwise answer *not equal*.

The state after Step 2 is

$$\frac{1}{\sqrt{(2k)!}} \sum_{\sigma \in S_{2k}} |\sigma\rangle \sigma(|\phi\rangle \cdots |\phi\rangle |\psi\rangle \cdots |\psi\rangle) \tag{7}$$

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

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: Assume $|\langle \phi | \psi \rangle| < \delta$. 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, which is

$$\left\| \frac{1}{\sqrt{(2k)!}} \sum_{\sigma \in S_{2k}} \langle 0 | F^\dagger |\sigma\rangle \sigma(|\phi\rangle \cdots |\phi\rangle |\psi\rangle \cdots |\psi\rangle) \right\|^2 \tag{8}$$

$$= \left\| \frac{1}{(2k)!} \sum_{\sigma \in S_{2k}} \sigma(|\phi\rangle \cdots |\phi\rangle |\psi\rangle \cdots |\psi\rangle) \right\|^2 \tag{9}$$

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

$$\leq \frac{(k!)^2}{(2k)!} (1+\delta)^{2k} \tag{11}$$

$$\sim \sqrt{\pi k} \left(\frac{1+\delta}{2} \right)^{2k}. \tag{12}$$

Finally, we consider briefly the case of fingerprinting where Alice and Bob have a shared *quantum* key, consisting of $O(\log n)$ 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 [4] 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 $n/2$ (and n is divisible by 4).

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 $O(\log n)$ Bell states that requires fingerprints of length only $O(\log n)$ bits. See [4] for details (the results are partly based on results in [5,7]). 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 exists also a classical scheme with classical keys and fingerprints of length $O(\log n)$.

Acknowledgments

We would like to thank Andris Ambainis, Ashwin Nayak, and John Preskill 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.

[1] A. Ambainis. Communication complexity in a 3-computer model. *Algorithmica*, 16(3):298–301, 1996.

[2] A. Ambainis, A. Nayak, A. Ta-Shma, and U. Vazirani. Quantum dense coding and a lower bound for 1-way quantum finite automata. In *Proceedings of 31st ACM STOC*, pages 376–383, 1999. quant-ph/9804043.

[3] N. Alon and J. H. Spencer. *The Probabilistic Method.* Wiley-Interscience, 1992.

[4] G. Brassard, R. Cleve, and A. Tapp. The cost of exactly simulating quantum entanglement with classical communication. *Physical Review Letters*, 83(9):1874–1877, 1999. quant-ph/9901035.

[5] H. Buhrman, R. Cleve, and A. Wigderson. Quantum vs. classical communication and computation. In *Proceedings of 30th ACM STOC*, pages 63–68, 1998. quant-ph/9802040.

[6] L. Babai and P. G. Kimmel. Randomized simultaneous messages: Solution of a problem of Yao in communication complexity. In *Proceedings of the 12th Annual IEEE Conference on Computational Complexity*, pages 239–246, 1997.

[7] P. Frankl and V. Rödl. Forbidden intersections. *Transactions of the American Mathematical Society*, 300(1):259–286, 1987.

[8] C. W. Helstrom. Detection theory and quantum mechanics. *Information and Control*, 10(1):254–291, 1967.

[9] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[10] J. Justesen, A class of constructive asymptotically good algebraic codes. *IEEE Trans. Inform. Theory*, 18:652–656, 1972.

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

[12] I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. In *Proceedings of 27th ACM STOC*, pages 596–605, 1995.

[13] F. MacWilliams and N. Sloane. *The Theory of Error-Correcting Codes.* North-Holland, 1977.

[14] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, 2000.

[15] I. Newman and M. Szegedy. Public vs. private coin flips in one round communication games. In *Proceedings of 28th ACM STOC*, pages 561–570, 1996.

[16] A. C-C. Yao. Some complexity questions related to distributive computing. In *Proceedings of 11th ACM STOC*, pages 209–213, 1979.