J Cryptol (2019) 32:601–634
https://doi.org/10.1007/s00145-019-09317-z

Journal of
CRYPTOLOGY

Key Establishment à la Merkle in a Quantum World

Gilles Brassard
Département IRO, Université de Montréal, Montreal, QC H3C 3J7, Canada
Canadian Institute for Advanced Research, Toronto, ON M5G 1M1, Canada
brassard@iro.umontreal.ca; http://www-labs.iro.umontreal.ca/~brassard/web/en/

Peter Høyer
Department of Computer Science, University of Calgary, Calgary, AB T2N 1N4, Canada
Canadian Institute for Advanced Research, Toronto, ON M5G 1M1, Canada
hoyer@ucalgary.ca

Kassem Kalach
ISARA Corporation, 560 Westmount Rd N, Waterloo, ON N2L 0A9, Canada
kassem.kalach@gmail.com

Marc Kaplan
VeriQloud, 13 rue Victor-Hugo, 92120 Montrouge, France
kaplan@VeriQloud.fr

Sophie Laplante
IRIF, Université Paris Diderot, Paris, France
laplante@irif.fr

Louis Salvail
Département IRO, Université de Montréal, Montreal, QC H3C 3J7, Canada
salvail@iro.umontreal.ca

Communicated by Stefan Wolf.

Received 5 December 2014 / Revised 3 March 2019
Online publication 8 April 2019

Abstract. In 1974, Ralph Merkle proposed the first unclassified protocol for secure communications over insecure channels. When legitimate communicating parties are willing to spend an amount of computational effort proportional to some parameter N, an eavesdropper cannot break into their communication without spending a time proportional to N^2, which is quadratically more than the legitimate effort. In a quantum world, however, Merkle's protocol is immediately broken by Grover's algorithm, but it is easily repaired if we are satisfied with a quantum protocol against which a quantum adversary needs to spend a time proportional to $N^{3/2}$ in order to break it. Can we do better? We give two new key establishment protocols in the spirit of Merkle's. The first

A preliminary version of this paper appeared in the Proceedings of Crypto 2011, Phil Rogaway (editor), but the results have been significantly strengthened in this archival version.
Kassem Kalach was at Université de Montréal when he worked on this research.

one, which requires the legitimate parties to have access to a quantum computer, resists any quantum adversary who is not willing to make an effort at least proportional to $N^{5/3}$, except with vanishing probability. Our second protocol is purely classical, yet it requires any quantum adversary to work asymptotically harder than the legitimate parties, again except with vanishing probability. In either case, security is proved for a typical run of the protocols: the probabilities are taken over the random (or quantum) choices made by the legitimate participants in order to establish their key as well as over the random (or quantum) choices made by the adversary who is trying to be privy to it.

Keywords. Merkle puzzles, Key establishment, Post-quantum cryptography, Quantum query complexity.

1. Introduction

While Ralph Merkle was delivering the 2005 International Association for Cryptologic Research (IACR) Distinguished Lecture at the CRYPTO annual conference in Santa Barbara, describing his original unpublished 1974 protocol [35] for public-key establishment (much simpler and more elegant than his subsequently published, yet better known, Merkle Puzzles [36]), one of us (Brassard) immediately realized that this protocol is totally insecure against an eavesdropper equipped with a quantum computer. The obvious question was: Can Merkle's idea be repaired and made secure again in our quantum world? The defining characteristics of Merkle's protocol are that (1) the legitimate parties communicate strictly through an authenticated classical channel on which eavesdropping is unrestricted and (2) a protocol is deemed to be *secure* if the cryptanalytic effort required of the eavesdropper to learn the key established by the legitimate parties grows super-linearly with the legitimate work.

Two of us (Brassard and Salvail [21]) partially repaired Merkle's idea in 2008 with a protocol in which the eavesdropper needs an amount of work in $\Omega\left(N^{3/2}\right)$ to obtain the key established by quantum legitimate parties whose amount of work is in $O(N)$. This was not quite as good as the work in $\Omega(N^2)$ required by a *classical* eavesdropper against Merkle's original protocol, but significantly better than the work in $O(N)$ sufficient for a *quantum* eavesdropper against the same protocol. Two main questions were left open in Ref. [21]:

1. Can the quadratic security possible in a classical world be restored in our quantum world?
2. Is any provable security possible at all if the legitimate parties are purely classical, yet the eavesdropper is endowed with a quantum computer?

At the CRYPTO 2011 conference, we gave two novel key establishment protocols to address these issues [18]. The current paper subsumes our earlier conference version and improves on it in various ways described below. In our first CRYPTO 2011 protocol, the legitimate parties use quantum computers and classical authenticated communication to establish a shared key after $O(N)$ expected queries to two random functions (which can be modelled with a single binary random oracle). We then gave a non-trivial quantum cryptanalytic attack, which enables a quantum eavesdropper to learn the key after $\Theta\left(N^{5/3}\right)$ queries to the functions. Furthermore, we proved that our attack is optimal (up to logarithmic factors) against that protocol. However, we initially focused on security

in the worst case, so that standard techniques in the theory of quantum lower bounds were sufficient. Unfortunately, this was of limited cryptographic relevance.

Our second Crypto 2011 protocol was *purely classical*, in the sense that the legitimate parties need only classical computation and classical communication to establish a key after $O(N)$ queries to similar random functions. We then gave a quantum cryptanalytic attack that requires $\Theta\left(N^{13/12}\right)$ queries to the functions. As unlikely as it may sound, this attack is optimal (again up to logarithmic factors) against that protocol, and therefore, it is not possible to break it with a quantum attack that uses an amount of resource linear in the legitimate effort. This was the first protocol ever proved secure (in the random oracle model) in the now thriving field of *post-quantum cryptography* [14,37]. However, our proof of security was also given only in the worst case, making it cryptographically unsatisfactory.

We now improve on the Crypto 2011 results in two directions. First, we simplify both protocols. Curiously, the simpler classical protocol is also more secure since $\Omega\left(N^{7/6}\right)$ quantum queries are required to break it, compared to $O\left(N^{13/12}\right)$ queries for the earlier protocol. Second, and more importantly, our proofs of security now hold for random instances rather than for the worst-case instance. It follows that the key obtained in a typical run of our protocols is secure, except with vanishing probability.

After a review of Merkle's original idea [35], its meltdown against a quantum eavesdropper and the obvious partial quantum solution [21] in Sect. 2, we describe our new protocols in Sects. 3 and 4, including quantum attacks against them in Sects. 3.1 and 4.1 and proofs of optimality for those attacks in Sects. 3.2 and 4.2. We then sketch an extension of these protocols to two families of more elaborate quantum and classical protocols in Sect. 5, but we postpone their detailed descriptions and proofs of security to a follow-up paper whose preliminary version is in Ref. [8]. Some practical aspects of our protocols are analysed in Sect. 6. As a technical tool needed in our proofs of lower bounds, we prove a new composition theorem of potential independent interest in Sect. 7. Finally, we conclude in Sect. 8 with a list of open problems in hope to improve our protocols in a variety of ways or prove intrinsic limits to such improvements.

2. Merkle's Original Protocol and How to Break and Partially Repair It with Quantum Computers

The first unclassified document ever written that pioneered public-key establishment and public-key cryptography was a class project proposal written in 1974 by Merkle when he was a student in Lance Hoffman's CS244 course on Computer Security at the University of California, Berkeley [35]. Hoffman rejected the proposal and Merkle dropped the course but "kept working on the idea" and eventually published it as one of the most seminal cryptographic papers in the second half of the twentieth century [36]. Merkle's protocol in his published paper was somewhat different from his original 1974 idea, but both share the property that they "force any enemy to expend an amount of work which increases as the square of the work required of the two [legitimate] communicants" [36]. It took 35 years before Boaz Barak and Mohammad Mahmoody-Ghidary proved that

this quadratic discrepancy between the legitimate and eavesdropping efforts is the best possible in a classical world [7] for provable security in the random oracle model.

In his IACR Distinguished Lecture,[1] which he delivered at the CRYPTO '05 Conference in Santa Barbara, Merkle described from memory his first solution to the problem of secure communications over insecure channels. As a wondrous coincidence, he unsuspectingly opened a box of old folders a mere three weeks after his Lecture and happily recovered his long-lost CS244 Project Proposal, together with comments handwritten by Hoffman [35]! To quote his original typewritten words:

```
Method 1:   Guessing.  Both sites guess at keywords.
            These guesses are one-way encrypted, and
            transmitted to the other site. If both sites
            should chance to guess at the same keyword,
            this fact will be discovered when the encryp-
            ted versions are compared, and this keyword
            will then be used to establish a communica-
            tions link.
Discussion: No, I am not joking.
```

In more modern terms, let f be a one-way permutation. In order to "one-way encrypt" x, as Merkle wrote in 1974, we assume that one can compute $f(x)$ in unit time for any given input x but that the only way to retrieve x given $f(x)$ is to try preimages and compute f on them until one is found that maps to $f(x)$. This is captured by the *random oracle* model. Accordingly, throughout this paper, with the exception of Sect. 6, efficiency is defined *solely* in terms of the number of queries to such oracles (there could be more than one). In the quantum case, these queries can be made in a superposition of inputs. We also assume throughout this paper (as did Merkle) that an authenticated channel is available between the legitimate communicants, although this channel offers no protection against eavesdropping.

The "keywords" guessed at by "both sites" are random points in the domain of f. They are "one-way encrypted" by applying f to them. If there are N^2 points in the domain of f, it suffices to guess $O(N)$ keywords at each site before it becomes overwhelmingly likely that "both sites should chance to guess at the same keyword", which becomes their shared key. An eavesdropper who listens to the entire conversation has no other way to obtain this key than to invert f on the revealed common encrypted keyword. In accordance with the oracle model, this can only be done by trying on average half the points in the domain of f before one is found that is mapped by f to the target value. This will require an expected number of queries to f in $\Omega(N^2)$, which is quadratic in the legitimate effort.

Shortly thereafter, Whitfield Diffie and Martin Hellman reinvented independently Merkle's notion of key establishment and discovered a celebrated method to achieve this goal, making the cryptanalytic effort *apparently* exponentially harder than the legitimate effort [25]. However, no proof is known that the Diffie–Hellman protocol is secure at all (even using elliptic curve cryptography) since it relies on the conjectured difficulty of extracting discrete logarithms, an assumption doomed to fail whenever quantum com-

[1] www.iacr.org/publications/dl/ann2005.html.

puters become available [40]. The same can be said of the subsequent (nowadays ubiquitous) RSA public-key cryptosystem [38]. In contrast, Merkle's approach offers provable quadratic security against any possible classical attack, under the sole assumption that f cannot be inverted by any other means than exhaustive search.

Next, we explain why Merkle's original proposal becomes completely insecure if the eavesdropper is capable of quantum computation. (Merkle's subsequently published "puzzles" [36] are equally insecure [21].) We then sketch our 2008 solution for a protocol that is not completely broken [21]. This is achieved by granting similar quantum computation capabilities to one of the legitimate communicating parties.

2.1. *Quantum Attack and Partial Remedy*

Let us now assume that function f can be computed quantum mechanically on a superposition of inputs. In this case, Merkle's original protocol is completely compromised by way of Grover's algorithm [29]. Indeed, this algorithm needs only query the function $O\left(\sqrt{N^2}\right) = O(N)$ times in order to invert it on any given point of its image, making the cryptanalytic task as easy (up to constant factors) as the legitimate key set-up process.[2]

To remedy the situation, Ref. [21] pioneered the idea of allowing the classically communicating parties to use quantum computers as well (actually, one of the parties may remain classical), and we increase the domain of f from N^2 to N^3 points. Instead of having both sites transmit one-way encrypted guesses to the other site, one site called Alice chooses N distinct random values $x_0, x_1, \ldots, x_{N-1}$ and transmits them, one-way encrypted by the application of f, to the other site called Bob. Let $Y = \{f(x_i) \mid 0 \leqslant i < N\}$ denote the set of encrypted keywords received by Bob, which becomes known to the eavesdropper. Now, Bob defines a Boolean function g on the same domain as f by

$$g(x) = \begin{cases} 1 & \text{if } f(x) \in Y \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Out of N^3 points in the domain of f, there are exactly $t = N$ solutions to the problem of finding an x so that $g(x) = 1$. It suffices for Bob to apply the BBHT generalization [15] of Grover's algorithm [29], which finds such an x after $O\left(\sqrt{N^3/t}\right) = O\left(\sqrt{N^2}\right) = O(N)$ queries to g and therefore to f. Bob sends back $f(x)$ to Alice, who knows the value of x because she was careful to keep her randomly chosen points. Therefore, $O(N)$ queries[3] to function f by Alice and Bob suffice for them to agree on key x.

The eavesdropper, on the other hand, is faced again with the need to invert f on a specific point of its image. Even with a quantum computer, this requires a number of

[2] If an unstructured search problem has t solutions among M candidates, Grover's algorithm [29], or more precisely its so-called BBHT generalization [15], can find one of the solutions after $O\left(\sqrt{M/t}\right)$ *expected* queries to a function that recognizes solutions among candidates. However, Theorem 4 of Ref. [19] implies that, whenever the number $t > 0$ is known, a solution can be found *with certainty* after $O\left(\sqrt{M/t}\right)$ queries to that function in the *worst case*. From now on, when we mention Grover's algorithm or BBHT, we really mean this improvement according to Ref. [19].

[3] If we cared about computational *efficiency* instead of only query complexity, Bob would sort the elements of Y in increasing order after receiving them from Alice. In this way, he can quickly determine, given any $y = f(x)$, whether or not $y \in Y$, which is needed to compute g. More on computational efficiency in Sect. 6.

queries to f proportional to the square root of the number of points in its domain [11], which is $\Omega\big(\sqrt{N^3}\big) = \Omega\big(N^{3/2}\big)$. This is more effort than what is required of the legitimate parties, yet less than quadratically so, as would have been possible in an all-classical world. Even though we have avoided the meltdown of Merkle's original approach, the introduction of quantum computers available to all sides seems to be to the advantage of the codebreakers. Can we remedy this situation? Furthermore, is any security possible at all against a quantum computer if both legitimate parties are restricted to being purely classical? We address these two questions in the rest of this paper.

3. Improved Quantum Key Establishment Protocol

The adjective *negligible* describes a function that decreases faster than the inverse of any polynomial. Formally, a function $\nu : \mathbb{N} \to \mathbb{R}$ is negligible if for any constant k, there exists N_k such that $\nu(N) < N^{-k}$ for all $N \geqslant N_k$. A weaker notion is that of *vanishing* function, which means that for any integer k, there exists N_k such that $\nu(N) < 1/k$ for all $N \geqslant N_k$, or, said otherwise, ν is $o(1)$. In cryptography, we usually strive for bad things to happen with negligible probability, such as the eavesdropper learning the key. Merkle's original work, however, was conceived in a way that a classical eavesdropper could not recover the secret key, except with vanishing probability, unless she made $O(N^2)$ queries to the oracle. Yet, a very lucky eavesdropper could recover the key with non-negligible probability $(1/N^2)$ with a single query to the oracle! Even though Merkle's protocol can be modified to make the eavesdropper's success probability negligible, rather than merely vanishing, provided we restrict her to at most $O(N^2/\log^2 N)$ queries [5], we shall be satisfied in this paper if the probability that a suitably bounded adversary can break our protocols is vanishing.

For any positive integer N, let $[N]$ denote the set of integers from 0 to $N-1$. For simplicity, we shall always take $N = 2^\ell$ to be a power of 2 and implicitly equate integer $i \in [N]$ with the ℓ-bit binary expansion of i seen as a bit string. This makes it possible to consider $[N]$ as a group under the bitwise exclusive-or operation, denoted "$\oplus$". In case one or both of i and j are the special symbol "$\star$" introduced later, we say that $i \oplus j$ is undefined, and therefore, it cannot be equal to w, regardless of what w is (not even $w = \star$). We describe our novel key establishment protocol assuming the existence of *two* random oracle functions $f : [N^3] \to [N^c]$ and $t : [N^3] \to [N^{c'}]$, where c and c' are constants discussed below. These oracles can be accessed in the usual quantum manner: for any $x \in [N^3]$ and $y \in [N^c]$, oracle f sends $|x, y\rangle$ to $|x, y \oplus f(x)\rangle$, and it sends superpositions of inputs to the corresponding superpositions of outputs; similarly for function t.

Constant c is chosen large enough so that f is one-to-one (there is no collision in the images of f), except with vanishing probability. An elementary calculation based on Boole's inequality (aka the union bound) shows that choosing $c > 6$ is sufficient. For simplicity, we shall henceforth disregard events that occur with vanishing probability, in particular the possibility that f not be one-to-one. Constant c' is chosen large enough to ensure that, except with vanishing probability, the function that maps unordered pairs $\{a, b\}$ of distinct elements to $t(a) \oplus t(b)$ is one-to-one. A similar calculation based again on Boole's inequality, using the fact that $\oplus$ maps uniformly distributed inputs to

uniformly distributed outputs, shows that $c' > 12$ is sufficient. Again, we shall henceforth assume that this property holds.

Notice that *a single binary random oracle* (which "implements" a random function from the integers to $\{0, 1\}$) could be used to define both functions f and t, provided we disregard logarithmic factors in our analyses, since $O(\log N)$ queries to the binary oracle would suffice to compute f or t on any given input. Indeed, to specify function f using a binary oracle, one needs only $N^3 \lg N^c$ bits from the binary oracle, where "$\lg$" denotes the binary logarithm, since each query $i \in [N^3]$ for f requires $\lg N^c$ queries to the binary oracle to construct the integer $f(i) \in [N^c]$. The situation is similar for function t. For this reason, it is understood hereinafter that all our results are implicitly stated "up to logarithmic factors". Furthermore, multiple function oracles can be encoded using a single binary oracle by prepending a fixed bit string to the beginning of each query. For instance, queries starting with "$0x$" and "$1x$" can be used to define $f(x)$ and $t(x)$, respectively.

Except in Sect. 6, where we care about computing *time*, the only resource that we consider in our analyses of efficiency and lower bounds is the number of queries made to these functions or, equivalently up to logarithmic factors, to the underlying binary random oracle.

Protocol 1. (Quantum vs. quantum)

1. Alice picks at random N distinct points $x_0, x_1, \ldots, x_{N-1}$ in the domain of f and transmits their encrypted values $y_i = f(x_i)$ to Bob. Let $X = \{x_i \mid 0 \leqslant i < N\}$ be the secret set of Alice and define $Y = \{y_i \mid 0 \leqslant i < N\}$. Note that Alice knows both X and Y, whereas Bob and the eavesdropper know only Y until they make their own queries to oracle f.
2. Bob finds the preimages x and x' of *two* distinct random elements in Y. For this purpose, he applies the BBHT generalization of Grover search *twice* on function g defined in Eq. (1), as reviewed in Sect. 2.1, using a small variation the second time to make sure that $x' \neq x$. If $f(x')$ was transmitted before $f(x)$ at Step 1, Bob swaps x and x'.
3. Bob sends back $w = t(x) \oplus t(x')$ to Alice.
4. Alice queries oracle t once on each element of X. No further query is required for her to find the two elements x_i and x_j in X such that $0 \leqslant i < j < N$ and $t(x_i) \oplus t(x_j) = w$.
5. The key shared by Alice and Bob is (x_i, x_j) for Alice and (x, x') for Bob, which is indeed the same under our assumptions on functions f and t.

All counted, Alice makes N classical queries to f in Step 1 and N classical queries to t in Step 4, whereas Bob makes $O(N)$ quantum queries to f in Step 2 and two classical queries to t in Step 3.

3.1. *Quantum Attack*

All the obvious (and some not so obvious) cryptanalytic attacks against this protocol, such as direct use of Grover's algorithm (or BBHT), or even more sophisticated attacks based on amplitude amplification [19], require the eavesdropper to query functions f

and t a total of $\Omega(N^2)$ times. However, a more powerful attack based on the paradigm of quantum walks in Markov chains [39] enables the eavesdropper to recover Alice and Bob's key with an expected $O\left(N^{5/3}\right)$ queries to f and $O\left(N^{2/3}\right)$ queries to t. This attack is reminiscent of Ambainis' quantum algorithm for element distinctness [3], which can find two elements i and j such that $e(i) = e(j)$ with $O\left(N^{2/3}\right)$ expected queries to any function e whose domain contains N elements, provided such elements exist.[4]

Ambainis' algorithm uses a quantum walk on the Johnson graph $J(N, r)$. This graph is an undirected graph in which each node contains an r-subset of $[N]$, meaning a subset of cardinality r for an appropriate value of r, and there is an edge between two nodes if and only if they differ by exactly one element. Intuitively, we may think of "walking" from one node to an adjacent node by dropping one element and replacing it by another. For the problem of element distinctness, the task is to find a 2-subset $\{i, j\}$ of $[N]$ such that $e(i) = e(j)$, provided one exists. The nodes that contain this subset are called *marked*. However, for a technical reason, our cryptanalytic task requires us to walk on a (modified) *Hamming graph* instead, in which the nodes contain lists rather than subsets, so that repetitions are allowed and the order in which items are listed matters.

Magniez, Nayak, Roland and Santha have proved a general theorem, showing that quantum search algorithms can be derived from a large class of classical Markov chains [34]. The cost of the resulting quantum algorithm can be written as a function of S, U and C. These are the cost of *setting up* the quantum register in a state that corresponds to the stationary distribution, *updating* it unitarily by walking from one node to an adjacent node and *checking* whether a node is marked in order to flip its phase if it is, respectively.

Theorem 1. ([34, Theorem 3]) *Let P be a reversible ergodic Markov chain with spectral gap $\delta > 0$. Then there is a quantum algorithm that finds a marked node, with high probability, provided there is at least one, at an expected cost in the order of*

$$\mathsf{S} + \frac{1}{\sqrt{\varepsilon}}\left(\frac{1}{\sqrt{\delta}}\mathsf{U} + \mathsf{C}\right),$$

where ε is the probability that a random node be marked.

Theorem 2. *There exists a quantum eavesdropping strategy that obtains the key established in Protocol 1 with $O\left(N^{5/3}\right)$ expected queries to f and $O\left(N^{2/3}\right)$ expected queries to t.*

Proof. Intuitively, we apply Ambainis' algorithm for element distinctness with two modifications: (1) instead of looking for $i \neq j$ such that $e(i) = e(j)$, we are looking for x and x' such that $t(x) \oplus t(x') = w$ and (2) instead of being able to get randomly chosen values in the image of e with a single query to oracle e per value, we need to get random elements of X by applying BBHT on the list Y and then query t on them, which

[4] There is no standard definition for the *element distinctness problem*. Ambainis uses the one above in Ref. [3] and our Definition 1 in Ref. [2], which is also used by others in Ref. [1]. This is of no significant importance because these two problems are easily seen to be computationally equivalent up to logarithmic factors in the query complexity model.

requires $O\left(\sqrt{N^3/N}\right) = O(N)$ queries to f and one query to t per element. The second modification explains why the number of queries to f, compared to $O\left(N^{2/3}\right)$ queries to e for element distinctness, is multiplied by $O(N)$. Hence, we need $O\left(N^{5/3}\right)$ queries to function f. To determine the number of queries required to function t, however, we have to delve deeper into the eavesdropping algorithm.

The composed structure of the problem prevents us from using a quantum walk on the Johnson graph, which was at the core of Ambainis' algorithm. Instead, we base the eavesdropping algorithm on a quantum walk on the Hamming graph $H(X, r)$, in which X is Alice's secret set and r is a number to be determined later. The nodes of the Hamming graph are labelled by ordered r-tuples of elements of X. There is an edge between two nodes when they differ on precisely one position. Said otherwise, the Hamming distance between their labels is 1. This graph has been used by Andrew Childs and Robin Kothari to study the quantum query complexity of minor-closed graph properties [24]. These authors have proved that the spectral gap δ of this graph is $\Omega(1/r)$. The quantum search algorithm on the Hamming graph also maintains a data structure at each node consisting of the image of each element of the node label under the random oracle t. In order to implement easily the update step of the quantum walk [34], we need to modify the Hamming graph by adding self-loops on all nodes, which does not change the spectral gap significantly [33]. Therefore, one can think of walking on the graph by replacing a randomly chosen element in the label of the current node by a randomly chosen element of X, thus leading to a self-loop with probability $1/N$.

We are looking for a marked node, which are those that contain two elements x and x' such that $t(x) \oplus t(x') = w$, where w is the value announced by Bob in Step 3 of the protocol. We use Theorem 1 on the modified Hamming graph, leading to a quantum search algorithm whose cost depends only on parameters S, U and C, as mentioned above. The set-up cost S corresponds to finding r not necessarily distinct random elements of X and then querying t on them. Technically, we need to obtain r independent equal superpositions of the elements of X by a unitary process, rather than r random elements, which is done by applying BBHT r times up to but excluding the final measurement. Since each use of BBHT requires $O(N)$ queries to f and one query to t, S consists of $O(rN)$ queries to f and r queries to t. The update cost U corresponds to finding one random element of X, which requires $O(N)$ queries to f, again by BBHT, and one query to t. Again, technically, the update consists of applying BBHT minus the final measurement to obtain an equal superposition of the elements of X, and then applying t on the result. The checking cost C requires us to decide whether there are distinct elements x and x' in the node such that $t(x) \oplus t(x') = w$, which can be done without any additional queries since all the relevant values of t are kept in the node. Finally, the probability ε for a random node to be marked is $\Omega\left(r^2/N^2\right)$. Putting it all together, the expected cryptanalytic cost is

$$\begin{aligned}
&\mathsf{S} + \tfrac{1}{\sqrt{\varepsilon}}\left(\tfrac{1}{\sqrt{\delta}}\mathsf{U} + \mathsf{C}\right) \\
&\quad = \mathsf{S} + O\left(\tfrac{N}{r}(\sqrt{r}\,\mathsf{U} + \mathsf{C})\right) \\
&\quad = \mathsf{S} + O\left(\tfrac{N}{\sqrt{r}}\,\mathsf{U}\right)
\end{aligned}$$

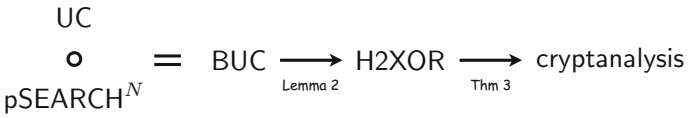

Fig. 1. Logical structure of the proof of Theorem 3 .

$$= O\left((rN \text{ queries to } f + r \text{ queries to } t) + \tfrac{N}{\sqrt{r}}\,(N \text{ queries to } f + \text{ one query to } t)\right)$$
$$= O\left(rN + N^2/\sqrt{r}\right) \text{ queries to } f \text{ **and** } O\left(r + N/\sqrt{r}\right) \text{ queries to } t\,.$$

To minimize the number of queries to f, we choose r so that $rN = N^2/\sqrt{r}$, which is $r = N^{2/3}$. It follows that a quantum eavesdropper is able to find the key with an expected $O\left(N^{5/3}\right)$ queries to f and $O\left(N^{2/3}\right)$ queries to t. □

3.2. *Lower Bound*

We prove in this section that the preceding quantum attack against our quantum protocol is optimal. This claim is formalized by the following theorem.

Theorem 3. *Any quantum eavesdropping strategy that recovers the key established in Protocol 1 requires a total of $\Omega\left(N^{5/3}\right)$ queries to functions f and t, except with vanishing probability. The vanishing probability is over a typical run of the protocol to establish a key, followed by the execution of an arbitrary quantum cryptanalytic algorithm to discover that key.*

Before we undertake the proof of this theorem, it is useful to summarize the task facing the adversary. After receiving a transcript of the protocol, which consists of $y_0, y_1, \ldots, y_{N-1}$ and w, where each $y_k = f(x_k)$ for an unknown x_k and $w = t(x_i) \oplus t(x_j)$ is a target value for some unknown i and j, she may make queries to functions f and t, after which she must determine x_i and x_j. Most of her queries to f produce irrelevant random values not appearing in the transcript. Among the relevant queries (hidden in random positions chosen by Alice), she still needs to solve the hard problem of finding two values that sum to w after function t is applied to them. Our proof that this is difficult for the adversary after a typical run of the protocol uses two intermediate problems: H2XOR, which is a "hidden" extension of the 2XOR problem (see Definitions 5 and 4) and a more structured problem called BUC (Bucketed Unique Collision, see Definition 3) for which we prove a worst-case lower bound for quantum query complexity. The "$\star$" symbol in the intermediate problems stand in for the irrelevant values of the random functions f and t.

More formally, the proof of this theorem consists of five steps, which are illustrated in Fig. 1.

1. We define the unique collision problem UC, the search problem pSEARCH and their composition BUC, which is related to the hardness of breaking our protocol;

2. We prove the hardness of BUC in the worst case (Corollary 1 of Theorem 9). For this purpose, we need a new composition theorem for the generalized adversary method, whose precise statement and technical proof are postponed to Sect. 7;
3. We define problems 2XOR and H2XOR, the latter being a hidden version of the former, which are more directly related to the hardness of breaking our protocol;
4. We prove an $\Omega\left(N^{5/3}\right)$ lower bound on the difficulty of solving H2XOR on random instances from the hardness of BUC in the worst case (Lemma 2);
5. We reduce H2XOR to the eavesdropping problem against our protocol. More precisely, we show that any attack on our key establishment protocol that would have a non-vanishing probability of success could be turned into an algorithm capable of solving H2XOR on random instances with twice the number of quantum queries (Theorem 3).

Notation 1. *For any set X, let $X^\star$ denote $X \cup \{\star\}$, where "$\star$" is some distinguished symbol that does not belong to X. Any element of X is called a non-$\star$ element of $X^\star$.*

Definition 1. (UC, COLL and ED Problems) Consider arbitrary integers N and $M \geqslant N$ and a function $e : [N] \to [M]$ so that there exist *exactly* two distinct elements i and j in $[N]$ for which $e(i) = e(j)$. Such a pair of elements is called a *collision*. The *unique collision* problem (UC) consists in finding these elements. Two related problems, called the *two-to-one collision problem* (COLL) and the *element distinctness* problem (ED), consist in finding a collision in a two-to-one function and deciding whether or not a given function is one-to-one, respectively.

Lemma 1. *The* UC *problem can be solved with $O(N^{2/3})$ quantum queries to function e, which is optimal in the worst case.*

Proof sketch. Andris Ambainis has given a quantum algorithm [3] capable of solving ED with $O(N^{2/3})$ quantum queries to function e; the same approach can be used to solve UC with the same efficiency. This algorithm for ED has been proved optimal in the worst case by Scott Aaronson and Yaoyun Shi [1] by reduction from their $\Omega(N^{1/3})$ lower bound for COLL. A variation on this reduction establishes an $\Omega(N^{2/3})$ lower bound on UC in the worst case because the restriction of any two-to-one function on a random subset of $\sqrt{N}$ points of its domain has constant probability of having a single collision (see Reduction 1.4 in Ref. [1]). Note that a lower bound on ED does not automatically yield the same lower bound on UC because the hard instances of ED could have been those that have either zero or multiple collisions if all we knew was the lower bound itself. □

Note that the lower bound on ED proved in Ref. [1] required $M \geqslant N^2$. A claim was made by Ambainis that this restriction is not necessary to establish the lower bound [2]. However, the proof given there is incomplete as it would apply only if the lower bound restricted to $M \geqslant N^2$ had been obtained with the polynomial method, which is not the case since it was obtained by a classical random reduction to COLL. The same applies to our lower bound on UC. Although the proof for the unrestricted case can be fixed [4], we do not need it for our purposes since we shall need to have $M \geqslant N^2$ for a different reason anyways.

Definition 2. (pSEARCH Problem) Let M and K be integers. Consider the set $P \subset ([M]^\star)^K$ of strings $(a_0, a_1, \ldots, a_{K-1})$ with the promise that exactly one value is non-$\star$. The problem pSEARCH : $P \to [M]$ of size K consists in finding this non-$\star$ value by making queries that take i as input and return a_i, $0 \leqslant i < K$. Unless stated otherwise, we shall use $K = N^2$.

Grover's algorithm [29] solves pSEARCH with $O\left(\sqrt{K}\right) = O(N)$ queries, and the first-ever lower bound on the power of quantum computing [11] shows that this is optimal in the worst case.

Definition 3. (Bucketed UC Problem) By composition, we define

$$\mathsf{BUC} = \mathsf{UC} \circ \mathsf{pSEARCH}^N.$$

Intuitively, an instance of BUC is obtained by "hiding" an instance of UC in "buckets" in which a single entry is non-$\star$. Formally, consider an instance $e : [N] \to [M]$ of the UC problem. An instance of BUC is given by a function $b : [N] \times [N^2] \to [M]^\star$. The domain of this function is composed of N buckets of size N^2 each, where $b(i, \cdot)$ corresponds to the ith bucket for $0 \leqslant i < N$. In bucket i, all values of the function are $\star$ except for *one single* $x_i \in [N^2]$ for which $b(i, x_i) = e(i)$.

$$b(i, j) = \begin{cases} e(i) & \text{if } j = x_i \\ \star & \text{otherwise}. \end{cases}$$

It follows from the definitions of e and b that there is a single pair of distinct α and β in the domain of b such that $b(\alpha) = b(\beta) \neq \star$, which is again called a *collision*. How difficult is it to find this collision given an oracle for b as only access to function e?

The query complexity of the BUC problem, as well as the above-mentioned lower bounds on the difficulty of solving the UC and pSEARCH problems, are stated and proved according to the usual complexity theoretic *worst-case* paradigm. This is clearly not what is needed for cryptographic applications. We shall remedy this situation shortly, starting with H2XOR in Lemma 2.

The BUC search problem is defined as the composition of UC with pSEARCH. Høyer, Lee and Špalek have proved a composition theorem for the quantum query complexity of such composed functions [31], later improved by Lee, Mittal, Reichard, Špalek and Szegedy [32]. Unfortunately, those theorems are not applicable in our case because they require the inner function to be Boolean, which pSEARCH is not.

Therefore, a more general composition theorem (Theorem 9) is needed, whose proof we postpone to Sect. 7 because of its technicality. We derive the statement we need here as Corollary 1 of Theorem 9, used with parameter $K = N^2$ (the size of the buckets), to obtain a lower bound of $\Omega(N^{5/3})$ on the worst-case query complexity of BUC.

The security of our key establishment protocol is based directly on neither the UC problem nor its bucketed version BUC, but rather on the 2XOR problem, or more precisely its "hidden" version H2XOR, both of which we now proceed to define.

Definition 4. (2XOR Problem) Consider arbitrary integers N and M, a nonzero *target* $w \in [M]$ and a one-to-one function $\xi : [N] \to [M]$ so that there exist exactly two distinct

elements i and j in $[N]$ for which $\xi(i) \oplus \xi(j) = w$. Such a pair of elements is called a *w-collision*. The 2XOR problem consists in finding these elements. The couple (w, ξ) is called an *N-M-instance* of this problem. A couple (w, ξ) such that ξ is not one-to-one or there are either no w-collisions, or more than one, will be called an *invalid* "instance".

It is elementary to adapt Ambainis' algorithm [3] for ED in order to solve the 2XOR problem with $O\left(N^{2/3}\right)$ quantum queries to function ξ, a fact that we do not actually need. Furthermore, it follows from Aaronson and Shi's matching lower bound [1] that 2XOR cannot be solved with fewer than $\Omega(N^{2/3})$ quantum queries in the worst case. Rather than proving this last statement (which we do not need either), we prove below the difficulty of uniformly distributed random instances of a problem more directly relevant to the analysis of our key establishment protocol.

Definition 5. (Hidden 2XOR Problem) Consider an N-M-instance (w, ξ) of the 2XOR problem and arbitrary distinct points $z_0, z_1, \ldots, z_{N-1}$ in $[N^3]$. Let function $h : [N^3] \to [N]^\star \times [M]^\star$ be defined as

$$h(x) = \begin{cases} (i, \xi(i)) & \text{if } x = z_i \text{ for some } i \in [N] \\ (\star, \star) & \text{otherwise .} \end{cases} \tag{2}$$

The unique pair of distinct points x and y in $[N^3]$ such that $\pi_2(h(x)) \oplus \pi_2(h(y)) = w$ is called a *w-collision*, where "π_k" denotes projection on the kth coordinate. The *hidden* 2XOR problem, or H2XOR for short, consists in finding this w-collision given w and oracle access to h; the couple (w, h) is called an *N-M-instance* of H2XOR.

Intuitively, given any N-M-instance (w, ξ) of the 2XOR problem, a corresponding instance of H2XOR consists in "hiding" the image of ξ among $N^3 - N$ uninformative $\star$ symbols in positions specified by the a priori unknown values of $z_0, z_1, \ldots, z_{N-1}$. The inherent difficulty of solving the 2XOR problem is therefore exacerbated by the difficulty of accessing the instance itself. The purpose of the intriguing first coordinate $[N]^\star$ in the image of h will become clear in the proof of Theorem 3.

We are now ready to prove that the H2XOR problem is difficult (in terms of query complexity) not only in the worst case but also on uniformly distributed random instances. (Note that the simpler 2XOR problem is also difficult on uniformly distributed random instances, but this is a fact that we do not need and therefore do not prove.)

Lemma 2. *Any quantum algorithm that attempts to solve the* H2XOR *problem with as few as $o\left(N^{5/3}\right)$ quantum queries in the worst case, for parameters N and $M \geqslant N^c$ with $c > 2$, will succeed with vanishing probability on a uniformly distributed random instance.*

Proof. Consider an arbitrary quantum algorithm $\mathcal{A}$ to solve the H2XOR problem. As usual in quantum query complexity, the algorithm is given in the form of a uniform family of quantum circuits, one for each value of N.[5] Each circuit alternates between

[5] For simplicity, we shall assume that the value of M is implicit given the value of N, so that we do not need to have a different circuit for each value of N and M. For instance, we could take $M = 2^{\lceil c \lg N \rceil}$, which would be the smallest power of 2 no smaller than N^c (note that we never said that c had to be an integer).

arbitrary oracle-independent unitary transformations and oracle queries. Let $p > 0$ and $q(N)$ be so that this algorithm succeeds with probability at least p on a random instance of H2XOR after at most $q(N)$ quantum queries. Even though $\mathcal{A}$ may only be designed to work on valid instances of H2XOR, nothing prevents us from running it on an invalid "instance", in particular if no w-collisions exist in h. In that case, $\mathcal{A}$ will obviously fail, but it will nevertheless do so after at most $q(N)$ queries.

Now, we proceed by a reduction of BUC to H2XOR. More specifically, we show how to transform an arbitrary instance of BUC into a uniformly distributed instance of H2XOR conditioned on an event whose probability is very close to 1/2, in such a way that a solution to the H2XOR instance provides a solution to the original BUC instance. It follows that the *worst-case* quantum query complexity lower bound for BUC proved in Corollary 1 of Theorem 9 translates to essentially the same quantum query complexity lower bound for H2XOR, but on uniformly distributed random instances.

For this purpose, consider an arbitrary instance $b : [N] \times [N^2] \to [M]^\star$ of the BUC problem, for the same parameters N and M, for which we wish to find the unique pair $\alpha = (\alpha_1, \alpha_2)$ and $\beta = (\beta_1, \beta_2)$ in $[N] \times [N^2]$ such that $b(\alpha) = b(\beta) \neq \star$. In order to randomize the instance, we choose two random permutations τ and σ, whose purpose is to shuffle the hidden values and their locations, respectively. Furthermore, we choose a random boolean function $\mathcal{L}$ used to add w to each non-$\star$ value of the BUC instance with probability 1/2. If w is added to one of the two elements in the BUC collision but not to the other (which happens with probability 1/2), this creates a w-collision in the resulting H2XOR instance. More formally, we choose a random permutation $\tau : [M] \to [M]$ and a random bijection $\sigma : [N^3] \to [N] \times [N^2]$. In order to transform b with bounded probability into a random instance (w, ξ) of H2XOR, we also choose a random nonzero target $w \in [M]$ and a random boolean function $\mathcal{L} : [N] \to \{0, 1\}$. Now, define function $h : [N^3] \to [N]^\star \times [M]^\star$ by

$$h(x) = \begin{cases} (\pi_1(\sigma(x)), \tau(b(\sigma(x)))) & \text{if } b(\sigma(x)) \neq \star \text{ and } \mathcal{L}(\pi_1(\sigma(x))) = 0 \\ (\pi_1(\sigma(x)), \tau(b(\sigma(x))) \oplus w) & \text{if } b(\sigma(x)) \neq \star \text{ and } \mathcal{L}(\pi_1(\sigma(x))) = 1 \\ (\star, \star) & \text{otherwise .} \end{cases}$$

Define $x = \sigma^{-1}(\alpha)$ and $y = \sigma^{-1}(\beta)$ so that $b(\sigma(x)) = b(\alpha) = b(\beta) = b(\sigma(y)) \neq \star$. Now, say that event SPLIT occurs if $\mathcal{L}(\alpha_1) \neq \mathcal{L}(\beta_1)$, which happens with probability 1/2. In case $\mathcal{L}(\alpha_1) = 0$ and $\mathcal{L}(\beta_1) = 1$, we have $h(x) = (\alpha_1, \tau(b(\alpha)))$ and $h(y) = (\beta_1, \tau(b(\beta)) \oplus w)$, and therefore, $\pi_2(h(x)) \oplus \pi_2(h(y)) = \tau(b(\alpha)) \oplus \tau(b(\beta)) \oplus w = w$ since $b(\alpha) = b(\beta)$. It follows that the pair x and y is a solution to instance (w, h) of H2XOR. The same conclusion holds if $\mathcal{L}(\alpha_1) = 1$ and $\mathcal{L}(\beta_1) = 0$. If we use algorithm $\mathcal{A}$ on this instance and it returns some pair (u, v) of distinct elements of $[N^3]$, it suffices to check whether or not $b(\sigma(u)) = b(\sigma(v)) \neq \star$, in which case we have solved the required instance of BUC with at most $q(N)$ queries to h, which is $O(q(N))$ queries to b since any quantum query on h can be computed via two quantum queries to b by use of standard reversible computing techniques [10].

However, it could be that (w, h) is an invalid "instance" of H2XOR. This could happen if SPLIT did not occur, in which case there is most likely no w-collisions in h (and if there is one, it is spurious). Even if SPLIT occurs, which guarantees the existence of at least one w-collision, (w, h) would be an invalid "instance" of H2XOR if and only

if there exist r and s in $[N] \times [N^2]$ such that $\tau(b(r)) \oplus \tau(b(s)) = w$, for one of two possible reasons: if $\mathcal{L}(\pi_1(r)) = \mathcal{L}(\pi_1(s))$, this creates a spurious w-collision in h, whereas otherwise the instance of the 2XOR problem hidden in function h is not one-to-one. The probability of any such occurrence is vanishing because $M \geqslant N^c$ for $c > 2$ and τ randomizes the values in the second coordinate of the range of h. Since it is impossible to determine efficiently if we have produced a valid instance of H2XOR, it could happen that we call algorithm $\mathcal{A}$ on an invalid (w, h). This has no other consequence than reduce the success probability by a constant factor, as we analyse in the next paragraph.

Define event VALID to mean that (w, h) is a valid instance of the H2XOR problem. The reader can verify that conditioned on both SPLIT and VALID this process generates a *uniformly distributed* instance of H2XOR. Furthermore, the probability of SPLIT is exactly $1/2$ and the probability that VALID fails is vanishing conditioned on SPLIT. Putting it all together, this algorithm produces a uniformly distributed instance of H2XOR with probability at least $1/2 - o(1)$, which by hypothesis is solved by $\mathcal{A}$ with at most $q(N)$ queries to h and correctness probability at least p. This yields an algorithm for BUC on an arbitrary instance b, using $O(q(N))$ queries to b, which is correct with non-vanishing probability at least $(1/2 - o(1))p$. Since solving BUC requires $\Omega\big(N^{5/3}\big)$ quantum queries in the worst case according to Corollary 1 (with parameter $K = N^2$), it follows that algorithm $\mathcal{A}$ also requires $\Omega\big(N^{5/3}\big)$ quantum queries in order to solve H2XOR with non-vanishing probability on a uniformly distributed random instance. □

We are now ready to return to the main theorem of this section, which concerns the cryptanalytic difficulty of breaking our key establishment protocol, and prove its security.

Proof of Theorem 3. Consider any eavesdropping strategy $\mathcal{A}$ that listens to the communication between Alice and Bob and tries to determine the key by querying functions f and t. In fact, there are no Alice and Bob at all! Instead, there is an instance (w, h) of H2XOR, hiding an instance (w, ξ) of 2XOR according to Eq. (2), which we want to solve by using unsuspecting $\mathcal{A}$ as a resource.

We start by supplying $\mathcal{A}$ with a completely fake "conversation" between "Alice" and "Bob": for sufficiently large c and c', we randomly choose N points $y_0, y_1, \ldots, y_{N-1}$ in $[N^c]$ and we pretend that Alice has sent the y's to Bob, who responded with the w from the instance of H2XOR that we want to solve. We also choose random functions $\hat{f} : [N^3] \to [N^c]$ and $\hat{t} : [N^3] \to [N^{c'}]$. Note that the selection of $\hat{f}$ and $\hat{t}$ may take a lot of *time*, but this does not count towards the number of *queries* that will be made to function h, and our lower bound on the search problem concerns *only* this number of queries. We could be tempted to choose randomly the values of $\hat{f}$ and $\hat{t}$ on the fly, whenever they are needed, but this is not an option for a quantum process because the values returned must be consistent whenever the same input is queried in different paths of the superposition.

Now, we wait for $\mathcal{A}$'s queries to f and t. When $\mathcal{A}$ queries $f(x)$ or $t(x)$ for some $x \in [N^3]$, we query $h(x)$. There are two possibilities.

- If $h(x) = (\star, \star)$, return $\hat{f}(x)$ and $\hat{t}(x)$ to $\mathcal{A}$ as value for $f(x)$ and $t(x)$, respectively. In this case, we say that x is *irrelevant*.

- Otherwise, let i be such that $h(x) = (i, \xi(i))$ and return y_i and $\xi(i)$ to $\mathcal{A}$ as value for $f(x)$ and $t(x)$, respectively. Intuitively, such an $f(x)$ corresponds to a relevant query in the simulated protocol.

The purpose of the first coordinate in the range of h now becomes clear. Whenever $h(x) \neq (\star, \star)$ and $f(x)$ is queried, the algorithm $\mathcal{A}$ should get one of the points y_0, y_1,..., y_{N-1} supplied to $\mathcal{A}$ at the beginning of this simulated cryptanalytic task. The same query $h(x)$ made subsequently should be answered consistently: $\mathcal{A}$ should get the same point. We might be tempted to choose an index on the fly and record it, but for the same reason as before with $\hat{f}$ and $\hat{t}$, it is not possible to keep track of these choices since the queries are made in superposition. Therefore, we provide the indices in the first coordinate of h.

Suppose $\mathcal{A}$ happily returns the pair (x, x') such that $t(x) \oplus t(x') = w$, which is what a successful eavesdropper is supposed to do. We return this pair, which is also a solution to our instance (w, h) of H2XOR, except with the vanishing probability that $\hat{t}(\tilde{x}) \oplus \hat{t}(\tilde{x}') = w$ for some irrelevant queries $\tilde{x}$ and $\tilde{x}'$ that $\mathcal{A}$ made to t.

To analyse the correctness of this reduction, we need to show that given a random instance (w, h) of H2XOR, it produces a random instance of the cryptanalytic task that $\mathcal{A}$ is purported to successfully solve. Notice that the functions f and t sampled by $\mathcal{A}$ are identical to random functions $\hat{f}$ and $\hat{t}$, except in N positions, where they are consistent with the y_i and corresponding $\xi(i)$, respectively. The values y_i are chosen at random; hence, f is random. However, since (w, h) is a *valid* instance of H2XOR, the random values of $\xi(i)$ hidden in h are all distinct; therefore, they are not fully independent. Nevertheless, the statistical distance between the resulting distribution on t and the uniform distribution is vanishing since the probability of a collision occurring in a subset of size N of indices of t would be vanishing under the uniform distribution.

Therefore, the environment provided by $\mathcal{A}$ in this simulation is the same as in the cryptanalytic context, except with vanishing probability. Since we disregard also the vanishing possibility that there might exist a spurious solution $t(\tilde{x}) \oplus t(\tilde{x}') = w$, on which $\mathcal{A}$ might happen, the reduction solves the search problem concerning h whenever $\mathcal{A}$ succeeds in finding the key. Notice finally that each (new) query made by $\mathcal{A}$ to either f or t translates to one query to h.

It follows that any successful cryptanalytic strategy that makes $o\left(N^{5/3}\right)$ total queries to f and t would solve the search problem with only $o\left(N^{5/3}\right)$ queries to h, which is impossible, except with vanishing probability, according to Lemma 2. This demonstrates the $\Omega\left(N^{5/3}\right)$ lower bound on the cryptanalytic difficulty of breaking our key establishment protocol on a typical run, again except with vanishing probability. □

4. Fully Classical Key Establishment Protocol

In this section, we revert to the original setting imagined by Merkle in the sense that Alice and Bob are now purely classical. However, we still allow full quantum power to the eavesdropper. Recall that Merkle's original protocols [35,36] are completely broken in this context [21]. Is it possible to restore *some* security in this highly adversarial (and unfair!) scenario? The following purely classical key establishment protocol, which is

inspired by our quantum protocol described in the previous section, provides a positive answer to this conundrum.

This time, random oracle functions f and t are defined on a smaller domain (N^2 instead of N^3) to compensate for the fact that classical Bob can no longer use the BBHT algorithm [15]. Specifically, $f : [N^2] \to [N^c]$ and $t : [N^2] \to [N^{c'}]$, with $c > 4$ and $c' > 8$ for reasons similar to those explained at the beginning of Sect. 3.

Protocol 2. (Classical vs quantum)

1. Alice picks at random N distinct points $x_0, x_1, \ldots, x_{N-1}$ in the domain of f and transmits their encrypted values $y_i = f(x_i)$ to Bob. Let X and Y denote $\{x_i \mid 0 \leqslant i < N\}$ and $\{y_i \mid 0 \leqslant i < N\}$, respectively.
2. Bob finds the preimages x and x' of two distinct random elements in Y. To find each one of them, he chooses random values in $[N^2]$ and applies f to them until one is found whose image is in Y. He is expected to succeed after $O(N)$ queries to f. If $f(x')$ was transmitted before $f(x)$ at Step 1, Bob swaps x and x'. Until now, this is almost identical to Merkle's original protocol, except for the fact that Bob needs to find two elements of X rather than one.
3. Bob sends back $w = t(x) \oplus t(x')$ to Alice.
4. Alice queries oracle t once on each element of X. No further query is required for her to find the two elements x_i and x_j in X such that $0 \leqslant i < j < N$ and $t(x_i) \oplus t(x_j) = w$.
5. The key shared by Alice and Bob is (x_i, x_j) for Alice and (x, x') for Bob, which is indeed the same.

All counted, Alice makes N queries to f in Step 1 and N queries to t in Step 4, whereas Bob makes $O(N)$ expected queries to f in Step 2 and two queries to t in Step 3. The total expected number of classical queries to f and t is therefore in $O(N)$ for both legitimate parties.

4.1. *Quantum Attack*

Theorem 4. *There exists a quantum eavesdropping strategy that obtains the key established in Protocol 2 with $O(N^{7/6})$ expected queries to f and $O(N^{2/3})$ expected queries to t.*

Proof. A quantum eavesdropper can set up a quantum walk very similar to the one explained in Sect. 3.1, except that now the domain is of size N^2 instead of N^3. The eavesdropper can find random elements of X from her knowledge of Y with an expected

$$O\left(\sqrt{N^2/N}\right) = O\left(\sqrt{N}\right)$$

queries to f per element of X. Therefore, the set-up cost S is in $O\left(r\sqrt{N}\right)$ queries to f and r queries to t, the update cost U is in $O\left(\sqrt{N}\right)$ queries to f and one query to t, and

the checking cost C vanishes. Furthermore, δ and ε are still in $\Omega(1/r)$ and $\Omega(r^2/N^2)$, respectively.

Putting it all together, the expected quantum cryptanalytic cost is

$$\begin{aligned}
&\mathsf{S} + O\left(\tfrac{N}{\sqrt{r}}\,\mathsf{U}\right) \\
&\quad = O\left((r\sqrt{N} \text{ queries to } f + r \text{ queries to } t) + \tfrac{N}{\sqrt{r}}(\sqrt{N} \text{ queries to } f + \text{one query to } t)\right) \\
&\quad = O\left(r\sqrt{N} + N^{3/2}/\sqrt{r}\right) \text{ queries to } f \ \textbf{and}\ O\left(r + N/\sqrt{r}\right) \text{ queries to } t\,.
\end{aligned}$$

To minimize the number of queries to f, we choose r so that $r\sqrt{N} = N^{3/2}/\sqrt{r}$, which is $r = N^{2/3}$ again. It follows that a quantum eavesdropper is able to find the key with an expected $O\left(N^{7/6}\right)$ queries to f and $O\left(N^{2/3}\right)$ queries to t. □

4.2. *Lower Bound*

The proof that it is not possible for the eavesdropper to find the key with fewer than a total of $\Omega\left(N^{7/6}\right)$ queries to f and t, except with vanishing probability, follows the same lines as the lower bound proof in Sect. 3.2. It is therefore possible for purely classical Alice and Bob to agree on a shared key after querying f and t an expected number of times in the order of N, whereas it is not possible, even for a *quantum* eavesdropper, to be privy to their secret with an effort in the same order, except with vanishing probability.

Theorem 5. *Any quantum eavesdropping strategy that recovers the key established in Protocol 2 requires a total of $\Omega\left(N^{7/6}\right)$ queries to functions f and t, except with vanishing probability. The vanishing probability is over a typical run of the protocol to establish a key, followed by the execution of an arbitrary quantum cryptanalytic algorithm to discover that key.*

Proof. The proof is similar to that of Theorem 3. The only difference is that Corollary 1 is applied with parameter $K = N$ (bucket size) rather than $K = N^2$. The proof then follows *mutatis mutandis*. □

5. Generalized Protocols

In Sects. 3 and 4, we presented a quantum and a classical protocol for key establishment over a classical channel. In both of them, Bob finds the preimages x and x' for f of two distinct elements sent by Alice, and he sends her back $t(x) \oplus t(x')$, which allows Alice to recover both x and x'. A natural generalization of these protocols is for Bob to find k preimages, for some constant $k \geqslant 2$, and send back to Alice the bitwise exclusive-or of the values of t applied to each one of them. Once Alice has recovered the k preimages found by Bob—we must increase the range of function t appropriately in order to ensure the uniqueness of the solution, except with vanishing probability—both Alice and Bob reorder them to reflect the order in which the images had been transmitted from Alice to Bob at the beginning of the protocol. The resulting k-tuple is the shared secret key.

This generalization leads to a sequence of quantum and classical protocols, denoted Q_k and C_k, respectively, with Q_2 and C_2 being Protocols 1 and 2 from the previous sections. These protocols still require Alice to make exactly N classical queries each to functions f and t, whereas Bob makes $O(kN)$ expected quantum or classical queries to f (depending on whether we are considering Q_k or C_k) and exactly k classical queries to t, which are simply $O(N)$ and $O(1)$ queries, respectively, because k is a constant.

Theorems 2 and 4 apply *mutatis mutandis* to show that quantum cryptanalytic attacks based on quantum walks on modified Hamming graphs succeed after $O(N^{1+k/(k+1)})$ expected queries to f and $O(N^{k/(k+1)})$ expected queries to t against Protocol Q_k, and $O(N^{1/2+k/(k+1)})$ expected queries to f and $O(N^{k/(k+1)})$ expected queries to t against Protocol C_k. For arbitrarily small ε, these attacks take a total number of queries in $O(N^{2-\varepsilon})$ and $O(N^{3/2-\varepsilon})$ against the quantum and classical protocols, respectively, provided k is sufficiently large.

The proof that these attacks are optimal against our generalized protocols is considerably more elaborate for the case $k > 2$ than when $k = 2$, corresponding to Theorem 3 for Q_2 and Theorem 5 for C_2. The problem kSUM, which is a natural generalization of 2XOR, was shown to be hard in the worst case by Belovs and Špalek [9]. However, there is no known way to prove a quantum lower bound on the difficulty of the kSUM problem *on uniformly distributed random instances* by a reduction from its difficulty in the worst case. Therefore, completely new tools had to be developed (with different co-authors) to prove the security of the generalized protocols, and hence conclude that Merkle's approach can be made essentially as secure in an all-quantum world as the original was in an all-classical world since our generalized protocols re-establish an arbitrarily close to quadratic security. This will be the topic of a follow-up paper whose preliminary version is in Ref. [8].

6. Practical Aspects of Our Protocols

In Sects. 3 to 5, we only counted the number of *queries* as a measure of complexity. In this section, we address the issue of whether the legitimate players have time-efficient strategies, as well as other "practical" considerations. It is important to understand that we do not claim that our protocols are actually practical, but only that some aspects of them can be made more realistic. After all, Merkle's original protocols [35,36] have never been deployed in real life, and this is obviously not because they are broken by Grover's algorithm [21]! Certainly, a very serious obstacle to the deployment of Merkle's protocols, which we do *not* address here, is the large amount of communication they may intrinsically require between the legitimate parties [30].

In any real implementation of our protocols, the random oracles would have to be replaced by quantum-resistant one-way functions, whose existence has not been established (and would require at the very least a proof that NP $\not\subseteq$ BPQ). Furthermore, even if we *had* such functions, the proofs of security for our protocols would not automatically carry through because of composability issues [22]. On the other hand, one might have objected to the notion of making queries in superposition to an oracle, whereas there are no issues about quantum computing a function on a superposition of inputs when it is specified by a quantum circuit. In any case, we shall assume in this section that functions

f and t from our protocols can be computed in constant time. If this is not the case, the time required by all parties includes the number of queries multiplied by the time it takes to compute these functions. An unfair case, which we do not consider here, may occur if these functions can be computed more efficiently on a quantum computer and if only the eavesdropper is endowed with one.

We now turn to the computational resources needed to implement our protocols. The first consideration is the time complexity of the legitimate players, be they classical or quantum. The second concerns the quantum protocols, where we address the issue of accessing quantum storage in superposition. All the key establishment protocols that we have presented share the following structure.

- Alice picks N points at random and sends the set Y of their images under function f to Bob.
- Bob searches for a set of preimages of a given size using either a classical or a quantum strategy, and sends it back to Alice, encoded.
- Alice recovers Bob's set, which becomes the key under a canonical ordering.

In the first step, Alice is only querying the oracle (or computing function f) and no post-processing is required. This can be done in $O(N)$ time.

For the second step, we showed that Bob needs only $O(N)$ expected queries per preimage, whether he is participating in the classical or quantum protocol. However, he may require an additional $\log N$ factor in terms of time because each query (whether or not in superposition) is followed by a binary search to check for membership in Y, as already mentioned in footnote 3 of Sect. 2.1. Thus, even though Bob needs only $O(N)$ queries, this translates into $O(N \log N)$ time. In the case of classical protocols, Bob can use universal hashing [23] to build a table for Y in $O(N)$ expected time, and then use it in constant expected time per search, so that his total expected time remains in $O(N)$. However, there is no obvious way to extend the use of universal hashing to the quantum protocols because all possible queries would be launched on the hash table in superposition, so that we would need good hashing performance in the worst case rather than in the expected sense. It turns out that a slight variation on our quantum protocols can guarantee a worst-case linear-time effort for Bob, as we now explain after a brief detour concerning a seldom recognized practical issue involving quantum memories.

Our quantum protocols require Bob to use a quantum memory to run the BBHT algorithm in his search for random elements of Alice's set X. Consider for instance the specific description of Step 2 in Protocol 1. It involves $O(N)$ Grover iterations. Each iteration involves a single call to g, as defined in Eq. (1), which itself involves one query to f (in a superposition of inputs), followed by a test of membership in Y of the output of f. This test requires the use of a memory of size N to hold Y, which must be accessible in a quantum superposition of its addresses (called a QRAM [27]) because f is queried in a superposition of all possible inputs (with non-uniform amplitudes in general) during each Grover iteration inside the BBHT algorithm. The use of such quantum memories has been a mostly unchallenged standard practice in quantum algorithmics at least since the 1997 paper of Ref. [20], but Daniel Bernstein objected as early as 2009 [13]. In the legitimate protocols presented here (but not in their cryptanalytic attacks), it suffices to have a memory that has to be loaded once with classical values (the elements of set Y), but that never needs to be updated once the quantum part of Bob's process—BBHT—

has been launched: this is in fact a QROM, which could be easier to implement than a QRAM. Nevertheless, Dominique Unruh pointed out that it may be unfair to count such quantum memory accesses at unit or even logarithmic cost in the memory size [41]. Be it as it may, quantum memories would likely be one of the most technologically challenging aspects to deploying our protocols, and therefore, it would be preferable if their need could be avoided [6].

We can modify our quantum protocols to remove any need for quantum memories, yet without compromising their security. We only sketch here the modifications that are needed for Protocol 1; the corresponding modifications for the generalized protocols outlined in Sect. 5 are identical, *mutatis mutandis*. Instead of having two functions $f : [N^3] \to [N^c]$ and $t : [N^3] \to [N^{c'}]$, we need $2N$ functions $f_i : [N^2] \to [N^c]$ and $t_i : [N^2] \to [N^{c'}]$, for $0 \leqslant i < N$. The first step of the protocol is the same, except that Alice defines each y_i as $f_i(x_i)$. In the second step, Bob chooses two indices $i < j$ at random in $[N]$. He uses the standard Grover algorithm (there is no need for BBHT anymore) to find preimages $x = x_i$ and $x' = x_j$ of y_i and y_j under f_i and f_j, respectively. This requires $O\left(\sqrt{N^2}\right) = O(N)$ Grover iterations *without any need for a quantum memory nor for an additional logarithmic factor in the time analysis*. The rest of the protocol is unchanged, except of course that Bob computes w as $t_i(x) \oplus t_j(x')$ and that Alice queries t_ℓ on each of her x_ℓ. Note that this modified protocol is more similar to Merkle's published "puzzles" [36], whereas the protocols we had described thus far are closer in spirit to Merkle's original unpublished idea [35].

The proof of security of the modified protocol is almost identical to the proof given in Sect. 3.2. The main difference is that we need to replace the H2XOR problem of Definition 5 by a new problem called B2XOR (for *Bucketed* 2XOR), which is to 2XOR what BUC is to UC. Formally,

$$\mathsf{B2XOR} = \mathsf{2XOR} \circ \mathsf{pSEARCH}^N.$$

Problems B2XOR and H2XOR are easily seen to be of equivalent query complexity, and the proof of Lemma 2 can be adapted to prove directly that B2XOR requires $\Omega(N^{5/3})$ queries on a uniformly distributed random instance from the worst-case lower bound on BUC given by Corollary 1. Furthermore, a uniformly distributed random instance of B2XOR can be solved with access to an adversary capable of breaking a typical instance of the modified key establishment protocol, following the lines of the proof of Theorem 3, which establishes the security of the modified protocol. We leave details to the reader. Note that the optimal attack against Protocol 1, given in the proof of Theorem 2, once adapted against the modified protocol, would still require the eavesdropper to make use of a quantum memory in order to perform quantum walks. Actually, the quantum walk paradigm [39] requires a quantum memory whose content is changed dynamically during the execution of the algorithm, which would be significantly more challenging from a technological point of view. However, following the usual paranoia in quantum cryptography, we are willing to grant the adversary unlimited technology, provided the laws of quantum theory are not violated.

Let us now turn our attention to the final process by which Alice recovers the key from the information she had kept and the information she has received from Bob. Let us first consider Protocols 1 and 2, although the modified Protocol 1 described above can be

handled in exactly the same way. Recall that Alice needs only N queries to function t since it suffices for her to obtain once each value of $t(x_i)$ and store them in a classical memory for future use. However, it may seem at first that she will need $\Omega(N^2)$ *time* to try a significant proportion of all the possible pairs among the N stored values of $t(x_i)$ before hitting upon two elements that exclusive-or to the value w received from Bob. This would obviously be intolerable. We now show that Alice can find the key time efficiently in these protocols.

Theorem 6. *Given w, Alice can use a classical algorithm to find two elements x and x' in X such that $t(x) \oplus t(x') = w$ in worst-case $O(N \log N)$ time or in expected $O(N)$ time.*

Proof. By querying t once on each element of X, Alice forms $Z = \{t(x) \oplus w \mid x \in X\}$ and she sorts it in $O(N \log N)$ time. Now, it suffices for her to try each value of $t(x')$, $x' \in X$, until one is found that belongs to Z. By definition of Z, there will be an $x \in X$ so that $t(x') = t(x) \oplus w$, which implies that $t(x) \oplus t(x') = w$ as required. Each of the (at most) N search operations is carried out in $O(\log N)$ time by virtue of using binary search, for a total of $O(N \log N)$ time in the worst case. Alternatively, Alice can use universal hashing [23] to build a table for Z in $O(N)$ expected time, and then search it in expected constant time per element of the form $t(x')$, $x' \in X$, for a total of $O(N)$ expected time. □

If we consider now the generalized protocols of Sect. 5, no classical algorithms are known that could handle Alice's last step efficiently whenever $k > 2$. However, a quantum Alice can do better, as shown in the following theorem, making Q_3 time-efficient as well.

Theorem 7. *Using a quantum strategy, Alice can find the elements x, x' and x'' in X such that $t(x) \oplus t(x') \oplus t(x'') = w$ in time $O(N \log N)$.*

Proof. By querying t once on each element of X, Alice forms $Z = \{t(x) \oplus w \mid x \in X\}$ and she sorts it in $O(N \log N)$ time. Then, she uses Grover's search algorithm to find a pair $(x', x'') \in X \times X$ such that $t(x') \oplus t(x'') \in Z$. It takes $O\left(\sqrt{N^2}\right) = O(N)$ Grover iterations to find this pair and each iteration takes $O(\log N)$ time by virtue of binary search in Z. Now, Alice can easily find the $x \in X$ such that $t(x') \oplus t(x'') = t(x) \oplus w$, which solves the problem since it follows that $t(x) \oplus t(x') \oplus t(x'') = w$, as desired. □

Unfortunately, the quantum algorithm in the proof of Theorem 7 requires the use of a quantum memory to hold Z. We do not know how to solve this problem otherwise. Table 1 summarizes the time separations that we get between the legitimate parties and the eavesdropper, although the adversary's lower bound for Q_3 requires a proof postponed to our follow-up paper whose preliminary version is in Ref. [8]. Recall that C_2 and Q_2 are Protocols 2 and 1, respectively. In each case, it is assumed that the adversary is capable of unrestricted quantum computation, including the use of a *dynamic* quantum memory, and that the legitimate parties agree on a shared key in $O(N)$—or at worst $O(N \log N)$—expected *time*. Only the last line in the table requires the use of a (static)

Table 1. Lower bounds on the time needed by quantum eavesdropping against various classical and quantum protocols when the legitimate parties establish a key in $O(N \log N)$ expected time .

Alice	Bob	Protocol	Adversary's lower bound
Classical	Classical	C_2	$\Omega(N^{7/6})$
Classical	Quantum	Q_2	$\Omega(N^{5/3})$
Quantum	Quantum	Q_3	$\Omega(N^{7/4})$

quantum memory on the part of one of the legitimate parties, provided the improvements suggested in this section are implemented.

7. A Composition Theorem for Quantum Query Complexity

To prove the hardness of breaking our protocols, we need to establish the worst-case hardness of a "hidden" extension BUC of the UC problem, whose quantum query complexity is known (see Lemma 1). We use the generalized adversary method, which we recall briefly below. This method is known to compose well (subject to some restrictions), and it is optimal, meaning that (up to a factor of 2) the optimal quantum query complexity is equal to the optimal adversary bound, for any function [31,32], a fact that we will use to our advantage to establish our result.

We briefly review the adversary method. Suppose we want to determine the quantum query complexity of a problem F. First, we assign weights to pairs of inputs in order to bring out how hard it is (in terms of number of queries) to distinguish these inputs from one another. The adversary lower bound is the worst ratio of the spectral norm of this matrix, which measures the overall progress necessary in order for the algorithm to be correct, to the spectral norms of associated matrices, which measure the maximum amount of progress that can be achieved by making a single query. For this purpose, we introduce the matrices D_q defined as follows:

$$D_q[x, y] = \begin{cases} 0 & \text{if } x_q = y_q \\ 1 & \text{otherwise.} \end{cases}$$

Definition 6. Fix a function $\mathsf{F} : S \to T$. A symmetric matrix $\Gamma : S \times S \to \mathbb{R}$ is an *adversary matrix* for F provided $\Gamma[x, y] = 0$ whenever $\mathsf{F}(x) = \mathsf{F}(y)$. The adversary bound of F using Γ is

$$\mathsf{ADV}^{\pm}(\mathsf{F}; \Gamma) = \min_q \frac{\|\Gamma\|}{\|\Gamma \bullet D_q\|},$$

where $\bullet$ denotes entrywise (or Hadamard) product, and $\|A\|$ denotes the spectral norm of A (which is equal to its largest eigenvalue). The adversary bound $\mathsf{ADV}^{\pm}(\mathsf{F})$ is the maximum, over all adversary matrices Γ for F, of $\mathsf{ADV}^{\pm}(\mathsf{F}; \Gamma)$.

Theorem 8. ([31,32]) *For any function* F, $\mathsf{Q}(\mathsf{F}) \leqslant \mathsf{ADV}^{\pm}(\mathsf{F}) \leqslant 2\mathsf{Q}(\mathsf{F})$, *where* $\mathsf{Q}(\mathsf{F})$ *is the worst-case quantum query complexity of* F.

Since BUC is defined as the composition of UC and $\mathsf{pSEARCH}$, we would like to apply a composition theorem for the generalized adversary method, which would say that if a function $\mathsf{H} = \mathsf{F} \circ \mathsf{G}^N$, then $\mathsf{ADV}^{\pm}(\mathsf{H}) \geqslant \mathsf{ADV}^{\pm}(\mathsf{F})\,\mathsf{ADV}^{\pm}(\mathsf{G})$. Unfortunately, the composition theorems already known in the literature [31,32] require the inner function to be Boolean, which is not the case here for $\mathsf{pSEARCH}$. Since counter-examples can be found [32], we cannot hope to prove a fully general composition theorem in which the inner function would be an arbitrary function. Nevertheless, we prove here a composition theorem with $\mathsf{pSEARCH}$ as the inner function.

Theorem 9. *For any* $\mathsf{F} : A^N \to B$, *let* $\mathsf{pSEARCH} : P \to A$ *with* $P \subseteq (A^\star)^K$ *be as in Definition 2, and define* $\mathsf{H} = \mathsf{F} \circ \mathsf{pSEARCH}^N$. *Then*

$$\mathsf{ADV}^{\pm}(\mathsf{H}) \geqslant \frac{2}{\pi}\,\mathsf{ADV}^{\pm}(\mathsf{F})\,\mathsf{ADV}^{\pm}(\mathsf{pSEARCH})\,.$$

The inner function can be slightly more general than $\mathsf{pSEARCH}$. For example, it could be that the element we search for is hidden in several places. The proof also goes through if the instances of $\mathsf{pSEARCH}$ operate over distinct domains $(A_i^\star)^{K_i}$. We leave for further research the extent to which our theorem can be generalized and proceed to prove it as stated.

Before we present the proof, we derive a corollary that is useful for our purpose, which applies for $\mathsf{BUC} = \mathsf{UC} \circ \mathsf{pSEARCH}^N$.

Corollary 1. $\mathsf{Q}(\mathsf{BUC}) = \Omega(N^{2/3}K^{1/2})$.

Proof of Corollary 1. From Lemma 1 together with Theorem 8, we have $\mathsf{ADV}^{\pm}(\mathsf{UC}) = \Omega(N^{2/3})$. Furthermore, $\mathsf{ADV}^{\pm}(\mathsf{pSEARCH}) > K^{1/2}$ by Eq. (10) below. Theorem 9 implies that $\mathsf{ADV}^{\pm}(\mathsf{BUC}) = \Omega(N^{2/3}K^{1/2})$. Finally, by Theorem 8, $\mathsf{Q}(\mathsf{BUC}) = \Omega(N^{2/3}K^{1/2})$ as well. □

Proof of Theorem 9. We prove the theorem using only a few properties of $\mathsf{pSEARCH}$, which we describe below. In order to discriminate between the N instances of $\mathsf{pSEARCH}$, and to simplify notation, we write the inner functions as $\mathsf{G}_1, \ldots, \mathsf{G}_N : P \to A$ with $P \subseteq (A^\star)^K$, $|A| = M$ and $|P| = MK$. We use the fact that G_i is K-to-1 for all i. Without loss of generality, we assume that inputs are sorted according to the output value. We use two crucial properties of $\mathsf{pSEARCH}$. These follow from the definition of an adversary matrix (Definition 6) as well as symmetry properties of $\mathsf{pSEARCH}$.

1. An $MK \times MK$ optimal adversary matrix Γ_i for G_i can be written in block form with $M \times M$ blocks of size $K \times K$ indexed by pairs of outputs in which all off-diagonal blocks are identical. Written in this form, all M diagonal blocks are necessarily zero since it is an adversary matrix.

$$\Gamma_i = \begin{pmatrix} 0 & S_i & \cdots & S_i \\ S_i & 0 & \ddots & S_i \\ \vdots & \ddots & \ddots & \vdots \\ S_i & S_i & \cdots & 0 \end{pmatrix} \qquad D_q = \begin{pmatrix} \Delta'_q & \Delta_q & \cdots & \Delta_q \\ \Delta_q & \Delta'_q & \ddots & \Delta_q \\ \vdots & \ddots & \ddots & \vdots \\ \Delta_q & \Delta_q & \cdots & \Delta'_q \end{pmatrix}$$

Fig. 2. The matrices Γ_i and D_q are decomposed into blocks $\Gamma_i^{(\tilde{x}_i,\tilde{y}_i)}$ and $D_q^{(\tilde{x}_i,\tilde{y}_i)}$, respectively. Each block labelled $\tilde{x}_i, \tilde{y}_i$ contains inputs x_i (resp. y_i) that map to the same output value, that is, $\mathsf{G}_i(x_i) = \tilde{x}_i$ (resp. $\mathsf{G}_i(y_i) = \tilde{y}_i$) .

2. The $MK \times MK$ matrices D_q, with inputs sorted in the same way, are also composed of identical off-diagonal blocks Δ_q and Δ'_q on-diagonal blocks. Notice that this strongly depends on G_i, since the inputs are sorted by output value.

For any function F, consider $\mathsf{H} = \mathsf{F} \circ (\mathsf{G}_1, \ldots, \mathsf{G}_N)$. Denote by I_M and $\mathbb{1}_M$ the $M \times M$ identity matrix and all-one matrix, respectively. We show that for all adversary matrices Γ_i for G_i of the form $\Gamma_i = (\mathbb{1}_M - I_M) \otimes S_i$, where S_i is a $K \times K$ symmetric matrix,

$$\mathsf{ADV}^{\pm}(\mathsf{H}) \geqslant \mathsf{ADV}^{\pm}(\mathsf{F}) \min_{i \in [N]} \mathsf{ADV}^{\pm}(\mathsf{G}_i; \Gamma_i). \tag{3}$$

To prove this, we define an adversary matrix Γ_{H} for H and compute its spectrum. It suffices to compute the largest eigenvalues of Γ_{H} and $\Gamma_{\mathsf{H}} \bullet D_q$ to give our lower bound on $\mathsf{ADV}^{\pm}(\mathsf{H})$.

Let us introduce some notation that we will use throughout the proof. Inputs to H are written $x, y \in P^N$. Each $x \in P^N$ breaks into $x = (x_1, \ldots, x_N)$. The result of applying the inner functions to $x = (x_1, \ldots, x_N)$ is written $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_N) = (\mathsf{G}_1(x_1), \ldots, \mathsf{G}_N(x_N))$. Each $x_i \in P$, seen as an element of $(A^\star)^K$, also breaks down into its components, which we write $x_i = ((x_i)_1, \ldots, (x_i)_K)$, where each component $(x_i)_j$ is an element of $A^\star$.

The structure on Γ_i allows us to consider it as $M \times M$ blocks, each of size $K \times K$, as follows. Rows and columns of Γ_i, indexed by inputs of the form $x_i = (a_1, \ldots, a_K) \in P$, are sorted according to the value $\tilde{x}_i = \mathsf{G}_i(x_i)$. The submatrix $\Gamma_i^{(\tilde{x}_i,\tilde{y}_i)}$ is the restriction of Γ_i to the rows and columns such that $\mathsf{G}_i(x_i) = \tilde{x}_i$ and $\mathsf{G}_i(y_i) = \tilde{y}_i$. When $\Gamma_i = (\mathbb{1}_M - I_M) \otimes S_i$, the diagonal blocks are the all-zero matrix and the others are equal to the matrix S_i. See Fig. 2.

We define Γ_{H} on blocks labelled by $(\tilde{x}, \tilde{y}) \in A^N \times A^N$. The submatrix $\Gamma_{\mathsf{H}}^{(\tilde{x},\tilde{y})}$ is the restriction of Γ_{H} to the rows and columns indexed by $x = (x_1, \ldots, x_N)$, $y = (y_1, \ldots, y_N) \in P^N$ such that $(\mathsf{G}_1(x_1), \ldots, \mathsf{G}_N(x_N)) = \tilde{x}$ and $(\mathsf{G}_1(y_1), \ldots, \mathsf{G}_N(y_N)) = \tilde{y}$:

$$\Gamma_{\mathsf{H}}^{(\tilde{x},\tilde{y})} = \Gamma_{\mathsf{F}}[\tilde{x}, \tilde{y}] \cdot \left(\bigotimes_{i=1}^{N} \overline{\Gamma}_i^{(\tilde{x}_i,\tilde{y}_i)} \right). \tag{4}$$

Here, Γ_{F} is an adversary matrix for F and instead of Γ_i, we have used the modified adversary matrices

$$\overline{\Gamma}_i = \Gamma_i + \|S_i\| I_{MK} ,$$

which add $\|S_i\|$ to the diagonal, to prevent zeroing out the block of H when $\tilde{x}_i$ equals $\tilde{y}_i$ on one of its components. The fundamental property of Γ_{H} is that its norm is the product of the norms of the matrices Γ_{F} and S_i.

Claim 1. *For the matrix Γ_{H} defined as above, $\|\Gamma_{\mathsf{H}}\| = \|\Gamma_{\mathsf{F}}\| \cdot \prod_{i=1}^{N} \|S_i\|$.*

We defer the proof of this claim and first see how it implies Eq. (3). Claim 1 gives us the norm of Γ_{H}, and it remains to compute $\max_\ell \|\Gamma_{\mathsf{H}} \bullet D_\ell\|$ (Definition 6). Let us turn to the matrix $\Gamma_{\mathsf{H}} \bullet D_\ell$ to see that it shares the structure of Γ_{H} so we can also apply Claim 1 to compute its norm. Recall that the domain of H is P^N, where $P \subseteq (A^\star)^K$. An index ℓ into an input x to H decomposes into $p \in [N]$, an index within x, and the index $q \in [K]$ within x_p seen as a vector in $(A^\star)^K$.

Claim 2. $\|\Gamma_{\mathsf{H}} \bullet D_\ell\| = \|\Gamma_{\mathsf{F}} \bullet D_p\| \cdot \|S_p \bullet \Delta_q\| \cdot \prod_{i \neq p} \|S_i\|$.

Proof of Claim 2. Restricting to the block labelled by $\tilde{x}$ and $\tilde{y}$, Ref. [31] shows that

$$(\Gamma_{\mathsf{H}} \bullet D_\ell)^{(\tilde{x},\tilde{y})} = (\Gamma_{\mathsf{F}} \bullet D_p)[\tilde{x}, \tilde{y}] \cdot \overline{(\Gamma_p \bullet D_q)}^{(\tilde{x}_p,\tilde{y}_p)} \otimes \left(\bigotimes_{i \neq p} \overline{\Gamma}_i^{(\tilde{x}_i,\tilde{y}_i)} \right). \tag{5}$$

Here we use the second property of pSEARCH: for each q, there exist matrices Δ_q and Δ'_q such that when restricted to blocks, $D_q = (\mathbb{1}_M - I_M) \otimes \Delta_q + I_M \Delta'_q$. Therefore, $\Gamma_p \bullet D_q$ has the same block structure as Γ_p and by Claim 1, we get the expression for $\|\Gamma_{\mathsf{H}} \bullet D_\ell\|$ given in Claim 2. □

Equation (3) follows from Claims 1 and 2.

$$\begin{aligned}
\mathsf{ADV}^{\pm}(\mathsf{H}; \Gamma_{\mathsf{H}}) &= \min_{p,q} \frac{\|\Gamma_{\mathsf{F}}\|}{\|\Gamma_{\mathsf{F}} \bullet D_p\|} \frac{\prod_{i=1}^{N} \|S_i\|}{\|S_p \bullet \Delta_q\| \cdot \prod_{i \neq p} \|S_i\|} \\
&= \min_{p,q} \frac{\|\Gamma_{\mathsf{F}}\|}{\|\Gamma_{\mathsf{F}} \bullet D_p\|} \frac{\|S_p\| \cdot \prod_{i \neq p} \|S_i\|}{\|S_p \bullet \Delta_q\| \cdot \prod_{i \neq p} \|S_i\|} \\
&= \min_{p,q} \frac{\|\Gamma_{\mathsf{F}}\|}{\|\Gamma_{\mathsf{F}} \bullet D_p\|} \frac{\|S_p\|}{\|S_p \bullet \Delta_q\|} \\
&\geqslant \min_{p} \left(\frac{\|\Gamma_{\mathsf{F}}\|}{\|\Gamma_{\mathsf{F}} \bullet D_p\|} \min_{q} \frac{\|S_p\|}{\|S_p \bullet \Delta_q\|} \right).
\end{aligned}$$

Using $\|\Gamma_i\| = (M-1)\|S_i\|$ and $\|\Gamma_i \bullet D_p\| = (M-1)\|S_i \bullet \Delta_p\|$, it follows that

$$\mathsf{ADV}^{\pm}(\mathsf{G}_p; \Gamma_p) = \min_q \frac{\|S_p\|}{\|S_p \bullet \Delta_q\|}, \tag{6}$$

and therefore,

$$\mathsf{ADV}^{\pm}(\mathsf{H}; \Gamma_{\mathsf{H}}) \geqslant \mathsf{ADV}^{\pm}(\mathsf{F}) \cdot \min_q \mathsf{ADV}^{\pm}(\mathsf{G}_q; \Gamma_q)\,.$$

Proof of Claim 1. We first prove $\|\Gamma_{\mathsf{H}}\| \leqslant \|\Gamma_{\mathsf{F}}\| \cdot \prod_i \|S_i\|$. The proof proceeds in four steps.

1. We define a set of vectors $\{\delta_{\alpha,c}\}$ in $\mathbb{C}^{(KM)^N}$.
2. We prove that they are eigenvectors of Γ_{H} and give the corresponding eigenvalues.
3. We show that we have defined all eigenvectors and eigenvalues of Γ_{H}.
4. We upper bound the eigenvalues in absolute value.

Similarly to the way we built up Γ_{H} from Γ_{F} and the Γ_i, we construct eigenvectors for Γ_{H} using the eigenvectors for Γ_{F} and the S_i as building blocks. We need some more notation before starting the proof. The spectrum of S_i is $\{(\delta_{i,j}, \lambda_{i,j})\}$ with eigenvalues $|\lambda_{i,1}| \geqslant \cdots \geqslant |\lambda_{i,K}|$. For $\tilde{x}_i, \tilde{y}_i \in A$, we use the following notation:

$$\lambda_{i,j}^{\tilde{x}_i \neq \tilde{y}_i} = \begin{cases} \lambda_{i,j} & \text{if } \tilde{x}_i \neq \tilde{y}_i \\ \|S_i\| & \text{otherwise.} \end{cases}$$

As we can see from the following eigenvalue equation, $\lambda_{i,j}^{\tilde{x}_i \neq \tilde{y}_i}$ is the eigenvalue of $\overline{\Gamma}_i^{(\tilde{x}_i, \tilde{y}_i)}$ associated with the vector $\delta_{i,j}$:

$$\begin{aligned} \overline{\Gamma}_i^{(\tilde{x}_i, \tilde{y}_i)} \delta_{i,j} &= \begin{cases} \lambda_{i,j}\delta_{i,j} & \text{if } \tilde{x}_i \neq \tilde{y}_i \\ \|S_i\|\delta_{i,j} & \text{otherwise} \end{cases} \\ &= \lambda_{i,j}^{\tilde{x}_i \neq \tilde{y}_i} \delta_{i,j}\,. \end{aligned} \tag{7}$$

Given a vector of indices $c = (c_1, \ldots, c_N)$, $c_i \in [K]$, we build up our eigenvectors for Γ_H by picking the c_ith eigenvector for the ith inner function (see Step 1). For $c = (c_1, \ldots, c_N)$, the $M^N \times M^N$ matrix A_c is defined by blocks

$$A_c[\tilde{x}, \tilde{y}] = \Gamma_{\mathsf{F}}[\tilde{x}, \tilde{y}] \cdot \prod_{i=1}^{N} \lambda_{i,c_i}^{\tilde{x}_i \neq \tilde{y}_i}$$

and we write its spectrum

$$\{(\alpha, \mu_{\alpha,c})\}\,.$$

Step 1: We are ready to define the eigenvectors $\delta_{\alpha,c}$ of Γ_{H}. We define the vectors $\delta_{\alpha,c}$ on the block $\delta_{\alpha,c}^{(\tilde{x})}$ of coordinates $x \in P^N$ such that $(\mathsf{G}_1(x_1), \ldots, \mathsf{G}_N(x_N)) = \tilde{x}$:

$$\delta_{\alpha,c}^{(\tilde{x})} = \alpha[\tilde{x}] \cdot \left(\bigotimes_{i=1}^{N} \delta_{i,c_i} \right) . \tag{8}$$

Notice that because of the structure of the Γ_i, it suffices for our purposes to build up the eigenvectors of Γ_{H} from the eigenvectors of the underlying S_i, which considerably simplifies the proof.

Step 2: We claim that the $\delta_{\alpha,c}$ are eigenvectors of Γ_{H} with corresponding eigenvalues $\mu_{\alpha,c}$. We want to calculate $\Gamma_{\mathsf{H}}\delta_{\alpha,c}$. We do this block by block. Fix $\tilde{x} \in A^N$. Using the eigenvalue equation (7), we get

$$\bigotimes_{i=1}^{N} \overline{\Gamma}_i^{(\tilde{x}_i,\tilde{y}_i)} \bigotimes_{i=1}^{N} \delta_{i,c_i} = \prod_{i=1}^{N} \lambda_{i,c_i}^{\tilde{x}_i \neq \tilde{y}_i} \bigotimes_{i=1}^{N} \delta_{i,c_i} . \tag{9}$$

Then, by Eqs. (4) and (8),

$$\begin{aligned}
(\Gamma_{\mathsf{H}}\delta_{\alpha,c})^{(\tilde{x})} &= \sum_{\tilde{y}} \left(\Gamma_{\mathsf{F}}[\tilde{x}, \tilde{y}] \cdot \bigotimes_{i} \overline{\Gamma}_i^{(\tilde{x}_i,\tilde{y}_i)} \right) \left(\alpha[\tilde{y}] \cdot \bigotimes_{i} \delta_{i,c_i} \right) \\
&= \sum_{\tilde{y}} \Gamma_{\mathsf{F}}[\tilde{x}, \tilde{y}]\alpha[\tilde{y}] \cdot \prod_{i} \lambda_{i,c_i}^{\tilde{x}_i \neq \tilde{y}_i} \cdot \bigotimes_{i} \delta_{i,c_i} \text{ (by Eq. 9)} \\
&= \sum_{\tilde{y}} A_c^{(\tilde{x},\tilde{y})} \alpha[\tilde{y}] \cdot \bigotimes_{i} \delta_{i,c_i} \\
&= \mu_{\alpha,c}\, \alpha[\tilde{x}] \cdot \bigotimes_{i} \delta_{i,c_i} \\
&= \mu_{\alpha,c}\, \delta_{\alpha,c} .
\end{aligned}$$

Step 3: We prove that the vectors $\delta_{\alpha,c}$ span $\mathbb{C}^{(KM)^N}$. There are K^N matrices A_c, and each one has M^N eigenvectors α. Therefore, $\{\delta_{\alpha,c}\}$ is a collection of $(KM)^N$ vectors. We now prove that they are orthogonal. Notice that

$$\begin{aligned}
\langle \delta_{\alpha,c}, \delta_{\alpha',c'} \rangle &= \sum_{\tilde{x}} \langle \delta_{\alpha,c}^{(\tilde{x})}, \delta_{\alpha',c'}^{(\tilde{x})} \rangle \\
&= \sum_{\tilde{x}} \left(\alpha[\tilde{x}]\alpha'[\tilde{x}] \cdot \prod_{i=1}^{N} \langle \delta_{i,c_i}, \delta_{i,c_i'} \rangle \right) \\
&= \langle \alpha, \alpha' \rangle \cdot \prod_{i=1}^{N} \langle \delta_{i,c_i}, \delta_{i,c_i'} \rangle .
\end{aligned}$$

If $\delta_{\alpha,c} \neq \delta_{\alpha',c'}$, it must be the case that either $c \neq c'$ or $\alpha \neq \alpha'$. Assume $c \neq c'$. Then for some i, $\delta_{i,c_i} \neq \delta_{i,c'_i}$ and since these vectors form an orthonormal basis of $\mathbb{C}^K$, we get $\langle \delta_{i,c_i}, \delta_{i,c'_i} \rangle = 0$. Now if $c = c'$, then $\alpha \neq \alpha'$. Again, these vectors form an orthonormal basis of $\mathbb{C}^{M^N}$ and we get $\langle \alpha, \alpha' \rangle = 0$.

Step 4: We prove by induction that the eigenvalues $\mu_{\alpha,c}$ of Γ_{H} are such that $|\mu_{\alpha,c}| \leqslant \|\Gamma_{\mathsf{F}}\| \cdot \prod_i \|S_i\|$ for all α and c. For $i \in [N+1]$ and $c \in [K]^N$, we define a family of matrices $A_c^{(i)}$ recursively as follows:

1. $A_c^{(0)} = \Gamma_{\mathsf{F}}$,
2. $A_c^{(i)}[\tilde{x}, \tilde{y}] = A_c^{(i-1)}[\tilde{x}, \tilde{y}] \cdot \lambda_{i,c_i}^{\tilde{x}_i \neq \tilde{y}_i}$.

By definition, $A_c^{(N)} = A_c$. We prove by induction that for each i,

$$\|A_c^{(i)}\| \leqslant \|\Gamma_{\mathsf{F}}\| \cdot \prod_{j=1}^{i} \|S_j\| .$$

Since $\mu_{\alpha,c}$ is an eigenvalue of A_c, this implies $|\mu_{\alpha,c}| \leqslant \|A_c\| \leqslant \|\Gamma_{\mathsf{F}}\| \cdot \prod_i \|S_i\|$.

Since $A_c^{(0)} = \Gamma_{\mathsf{F}}$, the base case is trivial. Assume that for some i, $\|A_c^{(i-1)}\| \leqslant \|\Gamma_{\mathsf{F}}\| \cdot \prod_{j=1}^{i-1} \|S_j\|$. By rearranging the rows and columns of $A_c^{(i-1)}$ as before, we can consider that it is formed of M^2 blocks with the following structure: the block labelled $(u, v) \in A \times A$ contains the entries $A_c^{(i-1)}[\tilde{x}, \tilde{y}]$ such that $\tilde{x}_i = u$ and $\tilde{y}_i = v$. Now, to form $A_c^{(i)}$, the diagonal blocks of $A_c^{(i-1)}$, labelled (u, u), are multiplied by $\|S_i\|$ and the others are multiplied by the same factor λ_{i,c_i}, which is at most $\|S_i\|$. We claim that under this operation, the norm of the matrix increases at most by a factor $\|S_i\|$.

Define $B = \frac{1}{|\lambda_{i,c_i}|} A_c^{(i)} - A_c^{(i-1)}$. This block diagonal matrix contains the diagonal blocks of $A_c^{(i-1)}$ multiplied by $\tau_i = \frac{1}{|\lambda_{i,c_i}|}\|S_i\| - 1$, while the other blocks are set to 0. In other words, B is a direct sum of operators acting on disjoint subspaces $E_1, \ldots, E_M$. It follows that

1. any eigenvalue of B is associated with an eigenvector whose support is in E_t for some t, and
2. for any vector v whose support is in E_t for some t, $\|Bv\| \leqslant \|\tau_i A_c^{(i-1)} v\|$.

This implies $\|B\| \leqslant \tau_i \|A_c^{(i-1)}\|$. Finally, writing $A_c^{(i)} = |\lambda_{i,c_i}|(A_c^{(i-1)} + B)$, we have

$$\begin{aligned} \|A_c^{(i)}\| &\leqslant |\lambda_{i,c_i}|(\|A_c^{(i-1)}\| + \|B\|) \\ &\leqslant |\lambda_{i,c_i}|(1 + |\tau_i|)\|A_c^{(i-1)}\|. \end{aligned}$$

Since λ_{i,c_i} is an eigenvalue of S_i, it is the case that $\tau_i \geqslant 0$, so $1 + |\tau_i| = \frac{1}{|\lambda_{i,c_i}|}\|S_i\|$. Finally,

$$\|A_c^{(i)}\| \leqslant \|S_i\| \cdot \|A_c^{(i-1)}\| .$$

The induction hypothesis allows us to conclude the proof of Step 4, which completes one direction in the proof of Claim 1.

We now prove the other direction: $\|\Gamma_{\mathsf{H}}\| \geqslant \|\Gamma_{\mathsf{F}}\| \cdot \prod_i \|S_i\|$. Taking $c = (1, \ldots, 1)$, we have $\|\Gamma_{\mathsf{H}}\| \geqslant \|A_c\|$. By definition, $A_c[\tilde{x}, \tilde{y}] = \Gamma_{\mathsf{F}}[\tilde{x}, \tilde{y}] \cdot \prod_i \|S_i\|$, which immediately implies that $\|\Gamma_{\mathsf{H}}\| \geqslant \|\Gamma_{\mathsf{F}}\| \cdot \prod_i \|S_i\|$. This completes the proof of Claim 1. □

To complete the proof of Theorem 9, we choose $S_i = \mathbb{1}_K$ and take $\Gamma_i = (\mathbb{1}_M - I_M) \otimes \mathbb{1}_K$ for the adversary matrix of $\mathsf{G}_i = \mathsf{pSEARCH}$, for each i. We verify that D_q has the necessary block structure. Indeed, for each output pair a, b of $\mathsf{pSEARCH}$, if $a \neq b$ then the block is all zero except in the row and column indexed by q, where it is 1, since the qth row corresponds to the input where a is hidden in position q and the qth column is the input where b is hidden in position q. Further, if $a = b$ then the block in D_q is 1 in column q and row q except in position (q, q) where it is zero. By direct computation, $\|S_i\| = K$ and $\|S_i \bullet \Delta_q\| = \sqrt{K-1}$. Using Definition 6 and Eq. (6) (with $\mathsf{G}_i = \mathsf{pSEARCH}$), it follows that

$$\begin{aligned} \mathsf{ADV}^{\pm}(\mathsf{pSEARCH}) &\geqslant \mathsf{ADV}^{\pm}(\mathsf{pSEARCH}; \Gamma_i) = \min_q \frac{\|S_i\|}{\|S_i \bullet \Delta_q\|} \\ &= \frac{K}{\sqrt{K-1}} > \sqrt{K}\,. \end{aligned} \tag{10}$$

By Theorem 8,

$$\mathsf{ADV}^{\pm}(\mathsf{pSEARCH})/2 \leqslant \mathsf{Q}(\mathsf{pSEARCH}) \leqslant \frac{\pi}{4}\sqrt{K}\,, \tag{11}$$

where Q denotes the quantum query complexity. Equations (10) and (11) imply that

$$\mathsf{ADV}^{\pm}(\mathsf{pSEARCH}; \Gamma_i) \geqslant \frac{2}{\pi}\,\mathsf{ADV}^{\pm}(\mathsf{pSEARCH})\,.$$

Theorem 9 now follows from Eq. (3). □

8. Conclusion and Open Questions

We live in a quantum world. Is this to the advantage of cryptographers or cryptanalysts [17]? The advantage to cryptographers is well established if they make use of quantum channels since quantum key distribution offers unconditional security [12], provided it is implemented faithfully [26]. However, the opposite seems to hold whenever cryptographers are limited to communicating over classical channels. Indeed, most of the cryptography currently deployed in attempts to protect information over the Internet fails completely in a quantum world [16]. The thriving field of post-quantum cryptography [14] aims at restoring at least some security for classical cryptography against a quantum adversary, and the race is on to find practical ways to do this [37]. In this paper, we gave a formal proof that this goal is achievable indeed, at least relative to a random oracle. For this, we showed how to restore some of the provable security of Merkle's seminal key

establishment protocol [35,36] in a quantum world, although we have not been capable of restoring the full quadratic advantage it enjoyed in an all-classical world. Not surprisingly, our protocol is more secure if the cryptographers are also endowed with quantum computing capabilities, but some security remains even if they are not. Along the way, we proved a composition theorem of potential independent interest in Sect. 7. We leave for further research the extent to which Theorem 9 can be generalized beyond the case of pSEARCH as the inner function.

We leave several other questions open for further research. Merkle's original protocol did not offer *negligible* probability of success against an adversary who would only be willing to invest an effort proportional to that used to establish the key. Indeed, after the cryptographers have invested an effort in the order of some security parameter N, a lucky eavesdropper could find the "secret" key with the same effort, albeit with probability $1/N$ (or even faster with a yet smaller probability). This is why we have deemed a protocol *secure* throughout this paper provided its probability of being broken by a resource-limited adversary is *vanishing* rather than negligible, as explained at the onset of Sect. 3. However, Merkle's original protocol *can* be modified to make the eavesdropper's success probability negligible, rather than merely vanishing, provided we restrict her to at most $O(N^2/\log^2 N)$ queries [5]. It would be interesting to see whether a similar approach can make our protocols provably quantum resistant if we required them to ensure secrecy except with negligible probability.

Our lower bounds prove that it is not possible for an eavesdropper to learn the key established by the protocol, except with vanishing probability, without querying the oracles significantly more than the legitimate parties. However, some *partial information* about the key leaks *even without querying the oracles at all*, and therefore, we do not achieve anything comparable to *semantic security* [28]. For instance, in Protocol 1, whenever the legitimate parties establish some key $k = (x, x')$, the eavesdropper learns the value of $t(x) \oplus t(x')$ since this is the revealed value of w. The question is whether or not the eavesdropper could learn any *useful* partial information about the key, whatever this could mean. Would there be any advantage in distilling $x \oplus x'$ as final key in Protocol 1 (or some other function of x and x'), rather than (x, x')? Alternatively, could the protocol be modified to probably deprive a resource-limited eavesdropper from *any* partial information?

We gave cryptanalytic attacks against our protocols that match the lower bounds on how difficult they are to attack successfully. However, those attacks proceed by a quantum walk in a modified Hamming graph, which requires the availability of a hypothetical QRAM [6,13,27]. In sharp contrast, our legitimate protocols (except those mentioned in Sect. 5) do not require such technological prowess provided they are modified according to Sect. 6. As an open question, can our protocols be attacked as efficiently without any need for QRAMs? Similarly, even if modified according to Sect. 6, protocol Q3 described in Sect. 5 seems to require at least availability of a QROM if it is to be time-efficient according to the proof of Theorem 7. Even though this may be easier to achieve than a QRAM, it would still represent a formidable technological challenge. Can protocol Q3 be implemented efficiently without any need for quantum memories? Even more interestingly, can generalized protocols C_k for $k > 2$ and Q_k for $k > 3$ be made time-efficient with or without need for fancy quantum memories?

Our protocols are defined in the random oracle model, which is why we are actually able to give formal proofs of security. As mentioned in Sect. 6, our random oracles would have to be replaced by quantum-resistant one-way functions (assuming they exist) before they can be deployed in real life. Unfortunately, our proofs of security would not carry automatically even if we had provable quantum-resistant one-way functions because of composability issues [22]. It would be much more interesting if we could prove the existence of quantum-secure classical key establishment protocols based only on the assumption that quantum-resistant one-way functions exist. This could be attempted either in proving composability properties of our protocols or in designing new protocols altogether.

The most important open question is to determine whether or not it is possible to restore Merkle's quadratic security in an all-quantum world, or perhaps even blast through that barrier since the proof of optimality for Merkle's protocol [7] only applies in the classical setting. Possibly a more practical consideration would be to determine the limit of provable security for a classical protocol based on random oracles against a quantum adversary. Partial answers to the questions raised in this paragraph are provided in our follow-up paper whose preliminary version is in Ref. [8], but those answers are severely limited if we take time efficiency into consideration.

Acknowledgements

We are grateful to Troy Lee, Frédéric Magniez, Mohammad Mahmoody-Ghidary, Miklos Santha and Robin Kothari for insightful discussions, to Krzysztof Pietrzak for pointing out the $O(N \log N)$-time algorithm that classical Alice can use in Protocols 1 and 2, and to Dominique Unruh for pointing out the "practical" difficulty (and possibly fundamental inefficiency) arising from the need to use quantum memories to implement some of our protocols. We are also indebted to an anonymous referee for pointing out that we don't necessarily have to be content with a vanishing probability that the eavesdropper learns the secret: protocols that would offer negligible probabilities are conceivable. The same referee also discovered a serious mistake in our original proof concerning the difficulty of solving H2XOR on a uniformly distributed random instance, which has now been fixed and became Lemma 2. G. B. is also grateful to Ralph Merkle for his most inspiring Distinguished Lecture at Crypto '05, which sparked this entire line of work, and to Harry Buhrman for hosting him at QuSoft during the final moments leading to the publication of this paper.

G. B. is supported in part by Canada's Natural Sciences and Engineering Research Council (Nserc), the Institut transdisciplinaire d'informatique quantique (Intriq), the Canada Research Chair Program and the Canadian Institute for Advanced Research (Cifar). P. H. is supported in part by Nserc, Cifar and the Canadian Network Centres of Excellence for Mathematics of Information Technology and Complex Systems (Mitacs). S. L. is supported in part by the projects EU FP7 QCS, EU CHIST-ERA DIQIP, EU QuantERA QuantAlgo, and the IRIF ICQ PICS Cooperation Project. L. S. is supported in part by an Nserc discovery grant and by Intriq.

References

[1] S. Aaronson, Y. Shi, Quantum lower bounds for the collision and the element distinctness problems. *Journal of the ACM* **51**(4), 595–605 (2004)

[2] A. Ambainis, Polynomial degree and lower bounds in quantum complexity: collision and element distinctness with small range. *Theory of Computing* **1**, 37–46 (2005)

[3] A. Ambainis, Quantum walk algorithm for element distinctness. *SIAM Journal on Computing* **37**, 210–239 (2007)

[4] A. Ambainis, Personal communication (2016)

[5] Anonymous Referee, Personal communication through the Editor, 26 October 2015

[6] S. Arunachalam, V. Gheorghiu, T. Jochym-O'Connor, M. Mosca and P. V. Srinivasan, On the robustness of bucket brigade quantum RAM. *New Journal of Physics* **17**(12), 123010 (2015)

[7] B. Barak, M. Mahmoody-Ghidary, Merkle puzzles are optimal—An $O(n^2)$–query attack on any key exchange from a random oracle, in *Advances in Cryptology—Proceedings of Crypto 2009* (Santa Barbara, California, 2009), pp. 374–390

[8] A. Belovs, G. Brassard, P. Høyer, M. Kaplan, S. Laplante, L. Salvail, Provably secure key establishment against quantum adversaries, in *Proceedings of 12th Conference on Theory of Quantum Computation, Communication and Cryptography (TQC)* (Paris, 2017), pp. 3:1–3:17. Open access available at http://drops.dagstuhl.de/opus/volltexte/2018/8581/pdf/LIPIcs-TQC-2017-3.pdf

[9] A. Belovs, R. Špalek, Adversary lower bound for the k-sum problem, in *Proceeding of 4th Annual ACM Conference on Innovations in Theoretical Computer Science (ITCS)* (Berkeley, California, 2013), pp. 323–328

[10] C. H. Bennett, Logical reversibility of computation. *IBM Journal of Research and Development* **17**(6), 525–532 (1973)

[11] C. H. Bennett, E. Bernstein, G. Brassard, U. V. Vazirani, Strengths and weaknesses of quantum computing. *SIAM Journal on Computing* **26**(5), 1510–1523 (1997)

[12] C. H. Bennett, G. Brassard, Quantum cryptography: Public key distribution and coin tossing, in *Proceedings of International Conference on Computers, Systems and Signal Processing* (Bangalore, India, 1984), pp. 175–179. Reprinted in *Theoretical Computer Science* **560-1**, 7–11 (2014)

[13] D. J. Bernstein, Cost analysis of hash collisions: Will quantum computers make SHARCS obsolete?, in *Proceedings of Workshop on Special-purpose Hardware for Attacking Cryptographic Systems (SHARCS'09)* (Lausanne, 2009), pp. 105–116. Proceedings available at http://www.hyperelliptic.org/tanja/SHARCS/record2.pdf

[14] D. J. Bernstein, T. Lange, Post-quantum cryptography. *Nature* **549**, 188–194 (2017)

[15] M. Boyer, G. Brassard, P. Høyer, A. Tapp, Tight bounds on quantum searching. *Fortschritte der Physik* **46**, 493–505 (1998)

[16] G. Brassard, Cryptography in a quantum world, in *42nd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)* (Springer, Harrachov, Czech Republic, 2016), pp. 3–16. Preliminary version available at arXiv:1510.04256 [quant-ph]

[17] G. Brassard, Was Edgar Allan Poe wrong after all? *Communications of the ACM* **62**(4), 132 (2019)

[18] G. Brassard, P. Høyer, K. Kalach, M. Kaplan, S. Laplante, L. Salvail, Merkle puzzles in a quantum world, in *Advances in Cryptology—Proceedings of Crypto 2011* (Santa Barbara, California, 2011), pp. 391–410

[19] G. Brassard, P. Høyer, M. Mosca, A. Tapp, Quantum amplitude amplification and estimation, in Samuel J. Lomonaco, Jr. editor, *Quantum Computation and Quantum Information*, *AMS Contemporary Mathematics*, **305**, 53–74 (2002)

[20] G. Brassard, P. Høyer, A. Tapp, Quantum algorithm for the collision problem (1997). arXiv:quant-ph/9705002

[21] G. Brassard, L. Salvail, Quantum Merkle puzzles, in *Proceedings of Second International Conference on Quantum, Nano, and Micro Technologies (ICQNM08)* (Sainte-Luce, Martinique, 2008), pp. 76–79

[22] R. Canetti, O. Goldreich, S. Halevi, The random oracle methodology, revisited (1998). https://eprint.iacr.org/1998/011
[23] L. Carter, and M. N. Wegman, Universal classes of hash functions. *Journal of Computer and System Sciences* **18**(2), 143–154 (1979)
[24] A. Childs, R. Kothari, Quantum query complexity of minor-closed graph properties, in *Proceedings of 28th Symposium on Theoretical Aspects of Computer Science (STACS)* (Dortmund, 2011), pp. 661–672
[25] W. Diffie, M. E. Hellman, New directions in cryptography. *IEEE Transactions on Information Theory* **22**(6), 644–654 (1976)
[26] I. Gerhardt, Q. Liu, A. Lamas-Linares, J. Skaar, C. Kurtsiefer, V. Makarov, Full-field implementation of a perfect eavesdropper on a quantum cryptography system. *Nature Communications* **2**, 349 (2011)
[27] V. Giovannetti, S. Lloyd, L. Maccone, Quantum random access memory. *Physical Review Letters* **100**(16), 160501 (2008)
[28] S. Goldwasser, S. Micali, Probabilistic encryption & How to play mental poker keeping secret all partial information, in *Proceedings of 14th Annual Symposium on Theory of Computing (STOC)* (San Francisco, California, 1982), pp. 365–377
[29] L. K. Grover, Quantum mechanics helps in searching for a needle in a haystack. *Physical Review Letters* **79**(2), 325–328 (1997)
[30] I. Haitner, N. Mazor, R. Oshman, O. Reingold, A. Yehudayoff, On the communication complexity of key-agreement protocols, in *Proceedings of 10th Innovations in Theoretical Computer Science Conference (ITCS)* (San Diego, California, 2019), paper no. 40. https://doi.org/10.4230/LIPIcs.ITCS.2019.40.
[31] P. Høyer, T. Lee, R. Špalek, Negative weights make adversaries stronger, in *Proceedings of 39th Annual Symposium on Theory of Computing (STOC)* (San Diego, California, 2007) pp. 526–535. The complete version can be found at arXiv:quant-ph/0611054
[32] T. Lee, R. Mittal, B. W. Reichardt, R. Špalek, M. Szegedy, Quantum query complexity of state conversion, in *Proceedings of the IEEE 52nd Annual Symposium on Foundations of Computer Science (FOCS)* (Palm Springs, California, 2011), pp. 344–353
[33] F. Magniez, Personal communication (2019)
[34] F. Magniez, A. Nayak, J. Roland, M. Santha, Search via quantum walk. *SIAM Journal on Computing* **40**(1), 142-164 (2011)
[35] R. Merkle, C.S. 244 Project Proposal (1974). Facsimile available at http://www.merkle.com/1974/FirstCS244projectProposal.pdf
[36] R. Merkle, Secure communications over insecure channels. *Communications of the ACM* **21**(4), 294–299 (1978)
[37] National Institute of Standards and Technology (NIST), Post-quantum cryptography standardization, https://csrc.nist.gov/projects/post-quantum-cryptography/post-quantum-cryptography-standardization
[38] R. L. Rivest, A. Shamir, L. Adleman, A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM* **21**(2), 120–126 (1978)
[39] M. Santha, Quantum walk based search algorithms, in *Proceedings of 5th Theory and Applications of Models of Computation (TAMC08)* (Xian, 2008), pp. 31–46
[40] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing* **26**, 1484–1509 (1997)
[41] D. Unruh, Objection raised during the question period when a preliminary version of this work was presented at the First Annual Conference on Quantum Cryptography (QCrypt), September 2011. Start at the 23rd minute of https://www.video.ethz.ch/conferences/2011/qcrypt/2011-09-12/5b98752b-7584-4ad0-b7fc-29aaf06371f9.html