Merkle Puzzles are Optimal — an $O(n^2)$-query attack on key exchange from a random oracle

Boaz Barak* Mohammad Mahmoody-Ghidary†

April 18, 2009

Abstract

We prove that every key exchange protocol in the random oracle model in which the honest users make at most n queries to the oracle can be broken by an adversary making $O(n^2)$ queries to the oracle. This improves on the previous $\tilde{\Omega}(n^6)$ query attack given by Impagliazzo and Rudich (STOC '89), and answers an open question posed by them. Our bound is optimal up to a constant factor since Merkle (CACM '78) gave a key exchange protocol that can easily be implemented in this model with n queries and cannot be broken by an adversary making $o(n^2)$ queries.

1 Introduction

In the 1970's Diffie, Hellman, and Merkle began to challenge the accepted wisdom that two parties cannot communicate confidentially over an open channel without first exchanging a secret key using some secure means. The first such protocol (at least in the open scientific community) was designed by Merkle in 1974 (although only published in 1978 [Mer]). Merkle's protocol allows two parties Alice and Bob to agree on a random number k that will not be known to an eavesdropping adversary Eve. It is described in Figure 1.

One problem with Merkle's protocol is that its security was only analyzed in the random oracle model which does not necessarily capture security when instantiated with a cryptographic one-way or hash function [CGH]. Recently, Biham, Goren and Ishai [BGI] took a step towards resolving this issue by providing a security analysis for Merkle's protocol under the concrete complexity assumption of existence of exponentially hard one-way functions. In particular, they proved that assuming there exist a one-way function that cannot be inverted with probability more than $2^{-\alpha n}$ by adversaries running in time $2^{\alpha n}$ for $\alpha \geq 1/2 - \delta$, there is a key exchange protocol in which Alice and Bob run in time n but any adversary whose running time is at most $n^{2-10\delta}$ has $o(1)$ chance of finding the secret. But the most serious issue with Merkle's protocol is that it only provides a *quadratic* gap between the running time of the honest parties and the adversary. Fortunately, not too long after Merkle's work, Diffie and Hellman [DH] and later Rivest, Shamir, and Adleman [RSA] gave constructions for key exchange protocols that are conjectured to have *super-polynomial* (even subexponential) security. But because these and later protocols are based on certain algebraic computational problems, and so could perhaps be vulnerable to unforseen attacks using this algebraic structure, it remained an important question to show whether there exist key exchange protocols with superpolynomial security that use only a random oracle.[2] The

*Department of Computer Science, Princeton University, `boaz@cs.princeton.edu`. Supported by NSF grants CNS-0627526, CCF-0426582 and CCF-0832797, US-Israel BSF grant 2004288 and Packard and Sloan fellowships.

†Department of Computer Science, Princeton University, `mohammad@cs.princeton.edu`. Supported by NSF grants CNS-0627526, CCF-0426582 and CCF-0832797.

[1] Merkle described his protocol using "puzzles" that can be implemented via some ideal cryptographic primitive; we describe the protocol in the case that the puzzles are implemented by a random oracle. We remark that in Merkle's original protocol Bob will try different random queries $y_1, y_2, \ldots$ without sending them to Alice until he finds y_j such that $f(y_j) \in \{a_1, \ldots, a_{10n}\}$ and send j — the index of the "puzzle" a_j — to Alice. The Protocol of Figure 1 is a *symmetric* version of Merkle's protocol, and is similar to the protocol of [CM] in the bounded storage model; see also discussion in [BGI].

[2] This is not to be confused with some more recent works such as [BR], that combine the random oracle model with assumptions on the intractability of other problems such as factoring or the RSA problem to obtain more efficient cryptographic constructions.

Merkle's Key Exchange Protocol

Let n be the security parameter. All parties have access to oracle to a function $H : \{0,1\}^\ell \to \{0,1\}^\ell$ chosen at random, where $\ell \gg \log n$. The protocol operates as follows:

1. Alice chooses $10n$ random numbers $x_1, \ldots, x_{10n}$ in $[n^2]$ and sends $a_1, \ldots, a_{10n}$ to Bob where $a_i = H(x_i)$ (embed $[n^2]$ in $\{0,1\}^\ell$ in some canonical way).
2. Bob chooses $10n$ random numbers $y_1, \ldots, y_{10n}$ in $[n^2]$ and sends $b_1, \ldots, b_{10n}$ to Alice where $b_j = H(x_j)$.
3. With at least 0.9 probability, there will be at least one "collision" between Alice's and Bob's messages: a pair i, j such that $a_i = b_j$. Alice and Bob choose the lexicographically first such pair, and Alice sets $s_A = x_i$ as her secret, and Bob sets $s_B = y_j$ as his secret. If no collision occurred they will not choose any secret. Note that assuming $2^\ell \gg n^4$, H will be one to one on $[n^2]$ with very high probability and hence $H(x_i) = H(y_j)$ implies $x_i = y_j$.

To analyze the protocol one shows that the collision is distributed uniformly in $[n^2]$ and deduces that an adversary Eve that makes $o(n^2)$ queries to the oracle will find the secret with $o(1)$ probability.

Figure 1: Merkle's key exchange protocol [Mer].[1]

seminal paper of Impagliazzo and Rudich [IR] answered this question negatively by showing that every key exchange protocol using n queries in the random oracle model can be broken by an adversary asking $O(n^6 \log n)$ queries.[3] Since a random oracle is in particular a one-way function (with high probability), this implied that there is no construction of a key exchange protocol based on a one-way function with a proof of super-polynomial security that is of the standard black-box type (i.e., a proof that transforms an adversary breaking the protocol into an inversion algorithm for the one-way function that only uses the adversary and the function as black boxes). Indeed, that was the motivation behind their result.

Question and motivation. Impagliazzo and Rudich [IR, Section 8] mention as an open question (which they attribute to Merkle) to find out whether their attack can be improved to $O(n^2)$ queries (hence showing the optimality of Merkle's protocol in the random oracle model) or there exist key exchange protocols in the random oracle model with $\omega(n^2)$ security. Beyond just being a natural question, it also has some practical and theoretical motivations. The practical motivation is that protocols with sufficiently large polynomial gap could be secure enough in practice — e.g., a key exchange protocol taking 10^9 operations to run and $(10^9)^6 = 10^{54}$ operations to break could be good enough for many applications.[4] In fact, as was argued by [BGI], as technology improves and honest users can afford to run more operations, such polynomial gaps only become more useful. Thus if known algebraic key exchange protocols were broken, one might look to polynomial-security protocol such as Merkle's for an alternative. Another motivation is theoretical— Merkle's protocol has very limited interaction (consisting of one round in which both parties simultaneously broadcast a message) and in particular it implies a public key encryption scheme. It is natural to ask whether more interaction can help achieve some polynomial advantage over this simple protocol. A third, less direct motivation comes from quantum computing. In one scenario in which some algebraic key exchange protocols will be broken— the construction of practical quantum computers— Merkle's protocol will also fail to offer non-trivial security due to Grover's search algorithm [Gro]. Our results below suggest (though do not prove) that Merkle's protocol may be optimal in this setting also, and so there may not exist a fully classical key-exchange protocol based on a one-way function with a black-box proof of super-linear security for quantum adversaries. We note that using quantum communication there is an *information theoretically* secure key-exchange protocol [BBE], and moreover, very recently Brassard and

[3] More accurately, [IR] gave an $O(m^6 \log m)$-query attack where m is the maximum of the number of queries n and the number of communication rounds, though we believe their analysis could be improved to an $O(n^6 \log n)$-query attack. For the sake of simplicity, when discussing [IR]'s results we will assume that $m = n$, though for our result we do not this assumption.

[4] Of course, these numbers are just an example and in practical applications the constant terms will make an important difference. We note though that the above constants are not ruled out by [IR]'s attack, but are ruled out by our attack (taking number of operations to mean the number of calls to the oracle).

Salvail [BS] (independently observed by [BGI]) gave a quantum version of Merkle's protocol, showing that if Alice and Bob can use quantum computation (but classical communication), to obtain a key-exchange protocol with super-linear (i.e., $n^{3/2}$) security in the random oracle model against quantum adversaries.

Our result. In this work we answer the above question of [IR], by showing that every protocol in the random oracle model where Alice and Bob make n oracle queries can be broken with high probability by an adversary making $O(n^2)$ queries. That is, we prove the following:

Theorem 1.1. *Let Π be a two-party protocol in the random oracle model such that when executing Π the two parties Alice and Bob make at most n queries each, and their outputs are identical with probability at least ρ. Then for every $0 < \delta < 1$, there is an adversary Eve making $(\frac{16n}{\delta})^2$ queries to the oracle whose output agrees with Bob's output with probability at least $\rho - \delta$.*

Our result can be shown to rule out the existence of *black-box* constructions of a key exchange protocol with super-quadratic security from a one-way function, though we omit the details.[5] To the best of our knowledge, no better bound than the $\tilde{O}(n^6)$-query attack of [IR] was previously known even in this case, where one does not assume the one-way function is a random oracle (hence making the task of proving a negative result easier). We note that similarly to previous black-box separation results, our adversary can be implemented efficiently in a relativized world where $\mathbf{P} = \mathbf{NP}$, meaning that we also rule out the somewhat larger family of *relativizing* reductions as well.

Correction of error: A previous version of this manuscript [BMG] posted on the Arxiv claimed a different proof of the same result. However, we have found a bug in that proof— see Appendix A. In fact the current proof is quite different from the one claimed in [BMG]. In [BMG] we also claimed an extension of Theorem 1.1 to the case of protocols with an oracle to a *random permutation* (i.e., a random one-to-one function R from $\{0,1\}^*$ to $\{0,1\}^*$ such that $|R(x)| = |x|$ for every $x \in \{0,1\}^*$). We do not know of an extension of the current proof to this model, beyond the observation of [IR] that any m-query attack in the random oracle model translates into an $O(m^2)$-query attack in the random permutation model. Hence our results imply an $O(n^4)$-query attack in the latter model, improving on the previous $\tilde{O}(n^{12})$ attack of [IR].

We also note that shortly after we posted the manuscript [BMG], Sotakova [SOT] posted an independently obtained weaker result, showing that protocols with only one round of interaction (each party sends one message) and non-adaptive queries can achieve at most $O(n^2)$ security. In contrast, as in the work of [IR], in this paper we allow protocols where the parties' choice of queries is adaptive and they can use an arbitrary polynomial number of interaction rounds.[6] The one-round case seems to be simpler, and in particular the bug found in our previous proof does not apply to that case.

2 Our techniques

The main technical challenge in proving such a result is the issue of *dependence* between the executions of the two parties Alice and Bob in a key exchange protocol. At first sight, it may seem that a computationally unbounded attacker that monitors all communication between Alice and Bob will trivially be able to find out their shared key. But the presence of the random oracle allows Alice and Bob to correlate their executions even without communicating (which is indeed the reason that Merkle's protocol achieves non-trivial security). Dealing with such correlations is the cause of the technical complexity in both our work and the previous work of Impagliazzo and Rudich [IR]. We handle this issue in a different way than [IR]. On a very vague high level our approach can be viewed as using more information about the structure of these correlations than [IR] did. This allows us to analyze a more efficient attacking algorithm, that is more frugal with the number of queries

[5] To formalize the above statement, one needs to quantify what it means for a black-box reduction to reduce a primitive Y to a primitive X with T security. Since as key size increases, the running time of X will dominate all other costs, we believe the correct formalization is that the reduction works for every adversary that makes at most T queries to the primitive X. Under this formalization, the above statement follows immediately from our results combined with the well-known fact that a random-oracle is a one-way function. Similar formalization (identifying running time with number of queries to the primitive) is needed to define the question for relativizing reductions.

[6] In fact, because we count only the number of *oracle queries* made by the honest parties, we can even allow a super-polynomial number of rounds.

it uses than the attacker of [IR]. Below we provide a more detailed (though still high level) exposition of our technique and its relation to [IR]'s technique.

2.1 Comparison with [IR].

We now review [IR]'s attack and outline of analysis, and particularly the subtle issue of *dependence* between Alice and Bob that arises in both their work and ours. The main novelty of our work is the way we deal with this issue, which is different from the approach of [IR]. We believe that this review of [IR]'s analysis and the way it compares to ours can serve as a useful introduction to our actual proof. However, no result of this section is used in the later sections, and so the reader should feel free at any time to skip ahead to Sections 3 and 4 that contain our actual attack and its analysis.

Consider a protocol that consists of n rounds of interaction, where each party makes exactly one oracle query before sending its message. [IR] called protocols of this type "normal-form protocols" and gave an $\tilde{O}(n^3)$ attack against them (their final result was obtained by transforming every protocol into a normal-form protocol with a quadratic loss of efficiency). Even though without loss of generality the attacker Eve of a key exchange protocol can defer all of her computation till after the interaction between Alice and Bob is finished, it is conceptually simpler in both [IR]'s case and ours to think of the attacker Eve as running concurrently with Alice and Bob. In particular, the attacker Eve of [IR] performed the following operations after each round i of the protocol:

- If the round i is one in which Bob sent a message, then at this point Eve samples $1000n \log n$ random executions of Bob from the distribution $\mathcal{D}$ of Bob's executions that are consistent with the information that Eve has at that moment (communication transcript and previous oracle answers). That is, Eve samples a uniformly random tape for Bob and uniformly random query answers subject to being consistent with Eve's information. After each time that she samples an execution, Eve asks the oracle all the queries asked during this execution and records the answers. (Generally, the true answers will not be the same answers as the one Eve guessed when sampling the execution.)
- Similarly, if the round i is one in which Alice sent a message then Eve samples $1000n \log n$ executions of Alice and makes the corresponding queries.

Overall Eve will sample $\tilde{O}(n^2)$ executions making a total of $\tilde{O}(n^3)$ queries. It's not hard to see that as long as Eve learns all of the *intersection queries* (queries asked by both Alice and Bob during the execution) then she can recover the shared secret with high probability. Thus the bulk of [IR]'s analysis was devoted to showing the following statement, denoted below by **(*)**: *With probability at least* 0.9 *Eve never fails, where we say that Eve* fails *at round i if the query made in this round by, say, Alice was asked previously by Bob but not by Eve.*

2.2 The issue of independence

At first look, it may seem that one could easily prove **(*)**. Indeed, **(*)** will follow by showing that at any round i, the probability that Eve fails in round i *for the first time* is at most $1/(10n)$. Now all the communication between Alice and Bob is observed by Eve, and if no failure has yet happened then Eve has also observed all the intersection queries so far. Because the answers for non-intersection queries are completely random and independent from one another it seems that Alice has no more information about Bob than Eve does, and hence if the probability that Alice's query q was asked before by Bob is more than $1/(10n)$ then this query q has probability at least $1/(10n)$ to appear in each one of Eve's sampled executions of Bob. Since Eve makes $1000n \log n$ such samples, the probability that Eve misses q would be bounded by $(1 - \frac{1}{10n})^{1000n \log n} \ll 1/(10n)$.

When trying to make this intuition into a proof, the assumption that Eve has as much information about Bob as Alice does translates to the following statement: conditioned on Eve's information, the distributions of Alice's view and Bob's view are *independent* from one another.[7] Indeed, if this statement was true then the above paragraph could be easily translated into a proof that [IR]'s attacker is successful, and it wouldn't have been hard to optimize this attacker to achieve $O(n^2)$ queries. Alas, this statement is false. Intuitively the reason

[7] Readers familiar with the setting of communication complexity may note that this is analogous to the well known fact that conditioning on any transcript of a 2-party communication protocol results in a product distribution (i.e., combinatorial rectangle) over the inputs. However, things are different in the presence of a random oracle.

is the following: even the fact that Eve has not missed any intersection queries is some non-trivial information that Alice and Bob share and creates dependence between them.[8]

Impagliazzo and Rudich [IR] dealt with this issue by a "charging argument" (see also Remark 2.1 below), where they showed that such dependence can be charged in a certain way to one of the executions sampled by Eve, in a way that at most n samples can be charged at each round (and the rest of Eve's samples are distributed correctly as if the independence assumption was true). This argument inherently required sampling at least n executions (each of n queries) per round, hence resulting in an $\Omega(n^3)$ attack.

2.3 Our approach

We now describe our approach and how it differs from the previous proof of [IR]. The discussion below is somewhat high level and vague, and glosses over some important details. Again, the reader is welcome to skip ahead at any time to Section 3 that contains the full description of our attack, and does not depend on this section in any way.

Our attacking algorithm follows the same general outline, but has two important differences from the attacker of [IR]:

1. One *quantitative* difference is that while our attacker Eve also computes a distribution $\mathcal{D}$ of possible executions of Alice and Bob conditioned on her knowledge, she does *not* sample from $\mathcal{D}$ full executions and then ask the arising queries. Rather, she computes whether there is any *heavy query*— a string $q \in \{0,1\}^*$ that has probability more than, say, $1/(100n)$ of being queried in $\mathcal{D}$— and makes only such heavy queries.

   Intuitively, since Alice and Bob make at most $2n$ queries, the total expected number of heavy queries (and hence the query complexity of Eve) is bounded by $O(n^2)$. The actual analysis is more involved since the distribution $\mathcal{D}$ keeps changing as Eve learns more information through the messages she observes and query answers she receives. We omit the details in this high-level overview.

2. The *qualitative* difference between the two attackers is that we do not consider the same distribution $\mathcal{D}$ that was considered by [IR]. Their attacker to some extent "pretended" that the conditional distributions of Alice and Bob are independent from one another. In contrast, we define our distribution $\mathcal{D}$ to be the *real* distribution of Alice and Bob, where there could be dependencies between them. Thus to sample from our distribution $\mathcal{D}$ one would need to sample a *pair* of executions of Alice and Bob (random tapes and oracle answers) that are *jointly consistent* with one another and Eve's current knowledge. Another (less important) point is that the distribution $\mathcal{D}$ computed by Eve at each point in time will be conditioned not only on Eve's knowledge so far, but also on the event that she has not failed until this point.

The main challenge in the analysis is to prove that the attack is *successful*, that is that the statement **(*)** above holds, and in particular that the probability of failure at each round (or more generally, at each query of Alice or Bob) is bounded by, say, $1/(10n)$. Once again, things would have been easy if we knew that the distribution $\mathcal{D}$ of the possible executions of Alice and Bob conditioned on Eve's knowledge (and not having failed so far) is a *product distribution*, and hence Alice has no more information on Bob than Eve has. While this is not generally true, we show that in our attack this distribution is *close to being a product distribution*, in a precise sense we define below.

At any point in the execution, fix Eve's current information about the system and define a bipartite graph G whose left-side vertices correspond to possible executions of Alice that are consistent with Eve's information and right-side vertices correspond to possible executions of Bob consistent with Eve's information. We put an edge between two executions A and B if they are consistent with one another and moreover if they do not represent an execution in which Eve *failed* prior to this point (i.e., there is no intersection query that is asked in both executions A and B but not by Eve). The distribution $\mathcal{D}$ that our attacker Eve considers can be thought of as choosing a random edge in the graph G. (Note that the graph G and the distribution $\mathcal{D}$ change at each

[8] As a simple example for such dependence consider a protocol where in the first round Alice chooses x to be either the string 0^n or 1^n at random, queries the oracle H at x and sends $y = H(x)$ to Bob. Bob then makes the query 1^n and gets $y' = H(1^n)$. Now even if Alice chose $x = 0^n$ and hence Alice and Bob have no intersection queries, Bob can find out the value of x just by observing that $y' \neq y$. Still, an attacker must ask a non-intersection query such as 1^n to know if $x = 0^n$ or $x = 1^n$.

point that Eve learns some new information about the system.) If G was the complete bipartite clique then $\mathcal{D}$ would be a product distribution. What we show is that G is *dense* in the sense that each vertex is connected to most of the vertices on the other side. We show that this implies that Alice's probability of hitting a query that Bob asked before is at most twice the probability that Eve does so if she chooses the most likely query based on her knowledge.

The bound on the degree is obtained by showing that G can be represented as a *disjointness graph*, where each vertex u is associated with a set $S(u)$ (from an arbitrarily large universe) and there is an edge between a left-side vertex u and a right-side vertex v if and only if $S(u) \cap S(v) = \emptyset$.[9] The definition of the graph G implies that $|S(u)| \leq n$ for all vertices u. The definition of our attacking algorithm implies that the distribution obtained by picking a random edge $\{u, v\}$ and outputting $S(u) \cup S(v)$ is *light*, in the sense that there is no element q in the universe that has probability more than $1/(10n)$ of being contained in a set chosen from this distribution. We show that these properties together imply that each vertex is connected to most of the vertices on the other side, and so G is close to being a complete bipartite graph.

Remark 2.1 (Comparison with [IR])**.** One can also phrase the analysis of [IR] in terms of a similar bipartite graph. Their argument involved fixing, say, Alice's execution which corresponds to fixing a left-side vertex u. As we noted above, if the degree of u is high (e.g., u is connected to most of the right side) then independence approximately holds and hence the probability that [IR]'s attacker fails at this point is less than $1/(10n)$. The crucial component of [IR]'s analysis was their observation that if the degree of u is low, then by taking a random vertex v on the right side and making all queries in the corresponding execution to v, one is likely to make progress in the sense that we learn a new query made in the execution corresponding to u. Now there are at most n new queries to learn, and hence if we sample much more than n queries then in most of them we're in the high degree case. This potential/charging argument inherently requires sampling *all* queries of the execution, rather than only the heavy ones, hence incurring a cost of at least n^2 queries *per round* or n^3 queries total. It also seems harder to generalize this argument to protocols that are not in normal form, which is the reason their attacker for general protocols required $\tilde{\Omega}(n^6)$ queries.

3 Our attacker

We consider a key exchange protocol Π in which Alice and Bob first toss coins r_A and r_B and then run Π using access to a random oracle H that is a random function from $\{0,1\}^\ell$ to $\{0,1\}^\ell$ for some $\ell \in \mathbb{N}$. We assume that the protocol proceeds in some finite number of rounds, and no party asks the same query twice. In round k, if k is odd then Alice makes some number of queries and sends a message to Bob (and then Eve asks some oracle queries), and if k is even then Bob makes some queries and sends a message to Alice (and then Eve asks some oracle queries). At the end of the protocol Alice obtains an output string s_A and Bob obtains an output string s_B. We assume that there is some constant $\rho > 0$ such that $\Pr[s_A = s_B] \geq \rho$, where the probability is over the coin tosses of Alice and Bob and the randomness of the oracle. We will establish Theorem 1.1 by proving that an attacker can make $O(n^2)$ queries to learn s_B with probability arbitrarily close to ρ.

In this section we describe an attacking algorithm that allows Eve to find a set of size $O(n^2)$ that contains all the queries asked by Alice and Bob in the random oracle model. This attack is analyzed in Section 4 to show that it is successful in finding all intersection queries and is efficient (i.e., will not ask more than $O(n^2)$ many queries). As was shown by Impagliazzo and Rudich, it not hard to use this set to obtain the actual secret.

3.1 Attacking algorithm

We start by showing that an attacker can find all the *intersection queries* (those asked by both Alice and Bob) with high probability. It turns out that this is the main step in showing that an attacker can find the secret with high probability (see Section 5.2 below).

Theorem 3.1. *Let Π be a key exchange protocol in the random oracle model in which Alice and Bob ask at most n oracle queries each. Then for every $0 < \delta < 1$ there is an adversary Eve who has access to the messages sent between Alice and Bob and asks at most $(\frac{13n}{\delta})^2$ number of queries such that Eve's queries contain all the intersection queries of Alice and Bob with probability at least $1 - \delta$.*

[9] The set $S(u)$ will correspond to the queries that are made in the execution corresponding to u but *not* made by Eve.

Letting $\epsilon = \delta/13$, our attack can be described in one sentence as follows:

> *As long as there exists a string q such that conditioned on Eve's current knowledge and assuming that no intersection query was missed so far, the probability that q was asked in the past (by either Alice or Bob) is at least ϵ/n, Eve makes the query q to the oracle.*

To describe the attack more formally, we need to introduce some notation. We fix n to be the number of oracle queries asked by Alice and Bob and assume without loss of generality that all the queries are of length $\ell = \ell(n)$ for some $\ell \in \mathbb{N}$. We will make the simplifying assumption that the protocol is in *normal form*— that is, at every round of the protocol Alice or Bob make exactly one query to the oracle (and hence there are $2n$ rounds). Later in Section 5.1 we will show how our analysis extends to protocols that are not of this form. Below and throughout the paper, we often identify a distribution $\mathcal{D}$ with a random variable distributed according to $\mathcal{D}$.

Executions and the distribution $\mathcal{EXEC}$. A (full) *execution* of Alice, Bob, and Eve can be described by a tuple (r_A, r_B, H) where r_A denotes Alice's random tape, r_B denotes Bob's random tape, and H is the random oracle (note that Eve is deterministic). We denote by $\mathcal{EXEC}$ the distribution over (full) executions that is obtained by running the algorithms for Alice, Bob and Eve with uniformly chosen random tapes and a random oracle. A *partial execution* is an execution truncated at a certain point in time (that is, the transcripts contain only the oracle answers for queries that are asked up to that point). For any partial execution we denote by M the sequence of messages sent between Alice and Bob till that moment, and denote by I the set of oracle query/answer pairs known to Eve. We define *Alice's view* in the execution to be the tuple $A = (r_A, H_A, M)$ where r_A are Alice's coins and H_A is the concatenation of oracle answers to Alice's queries. Similarly Bob's view is the tuple $B = (r_B, H_B, M)$. Below we will only consider Alice's and Bob's view conditioned on a fixed value of M and hence we drop M from these tuples and let $A = (r_A, H_A)$ and $B = (r_B, H_B)$.

The distribution $\mathcal{EXEC}(M, I)$. For $M = [m_1, \ldots, m_i]$ a sequence of i messages, and I a set of query/answer pairs, we denote by $\mathcal{EXEC}(M, I)$ the distribution over the views (A, B) of Alice and Bob in partial executions up to the point in the system in which the i^{th} message is sent (by Alice or Bob), where the transcript of messages equals M and the set of query/answer pairs that Eve learns equals I. For every (M, I) that have nonzero probability to occur in the protocol, the distribution $\mathcal{EXEC}(M, I)$ can be sampled by first sampling (r_A, r_B, H) at random conditioned on being consistent with (M, I) and then deriving from this tuple Alice's and Bob's views: $A = (r_A, H_A)$ and $B = (r_B, H_B)$.[10]

The event $\mathsf{Good}(M, I)$ and the distribution $\mathcal{GEXEC}(M, I)$. The event $\mathsf{Good}(M, I)$ is defined as the event over $\mathcal{EXEC}(M, I)$ that all the intersection queries asked by Alice and Bob during the partial execution are in I. More formally let $Q(A)$ (resp. $Q(B)$) be the set of queries asked by Alice (resp. Bob) which are specified by Alice's view A (resp. Bob's view B). Therefore $\mathsf{Good}(M, I)$ is the same as $Q(A) \cap Q(B) \subset Q(I)$ where $Q(I)$ is the set of queries of I (note that I is a set of query/answser *pairs*). We define the distribution $\mathcal{GEXEC}(M, I)$ to be the distribution $\mathcal{EXEC}(M, I)$ conditioned on $\mathsf{Good}(M, I)$.

Eve's algorithm. The attacker Eve's algorithm is specified as follows. It is parameterized by some constant $\epsilon > 0$ which we assume is smaller than $1/10$. At any point in the execution, if M is the sequence of messages Eve observed so far and I is the query/answer pairs she learned so far, Eve computes for every $q \in \{0, 1\}^{\ell}$ the probability p_q that q appears as a query in a random execution in $\mathcal{GEXEC}(M, I)$. If $p_q > \epsilon/n$ then Eve asks q from the oracle and adds q and its answer to I. (If there is more than one such q then Eve asks the lexicographically first one.) Eve continues in this way until there is no additional query she can ask, at which point she waits until she gets new information (i.e., observes a new message sent between Alice and Bob).

Note that Eve's algorithm above may ask much more than n^2 queries. However, we will show that the probability that Eve asks more than n^2/ϵ^2 queries is bounded by $O(\epsilon)$, and hence we can stop Eve after asking this many queries without changing significantly her success probability.

---

[10] Note that we can verify that the pair (M, I) has nonzero probability to occur in the protocol by simulating Eve's algorithm on the transcript M, checking that whenever Eve makes a query, this query is in I, in which case we feed Eve with the corresponding answer (and verifying at the end that there are no "extra" queries in I not asked by Eve). However in our attack the pair (M, I) will always be generated by running the actual protocol and so we won't need to run such checks.

Remark 3.2. The attacking algorithm above is not computationally efficient, as in general computing the probability distribution $\mathcal{GEXEC}(M, I)$ could be a hard problem since it involves "inverting" the algorithms of Alice and Bob to a certain extent. But because computing these probabilities is in $\#\mathbf{P}$, we can use known techniques (e.g., [BGP]) to approximate them with arbitrarily good precision using an $\mathbf{NP}$-oracle. In particular this means that our attacker (as was the case in previous works) is computationally efficient in a relativized world in which $\mathbf{P} = \mathbf{NP}$, and hence this result also rules out *relativizing* reductions from one-way functions to key exchange that achieve $\omega(n^2)$ security (when this is formalized as mentioned in Footnote 5, identifying running time with number of calls to the one-way function).

4 Analysis of attack: proof of Theorem 3.1

We now go over the proof of Theorem 3.1. For $i \in [2n]$, define the event Fail_i to be the event that the query made at the i^{th} round is an intersection query but is not contained in the set I of query/answer pairs known by Eve, and moreover that this is the first query satisfying this condition. Let the event $\mathsf{Fail} = \bigvee_i \mathsf{Fail}_i$ be the event that at some point an intersection query is missed by Eve, and let the event Long be that Eve makes more than n^2/ϵ^2 queries. By setting $\epsilon = \delta/13$ and stopping Eve after n^2/ϵ^2 queries, Theorem 3.1 immediately follows from the following two lemmas:

Lemma 4.1 (Attack is successful)**.** *For every i, $\Pr_{\mathcal{EXEC}}[\mathsf{Fail}_i] \le \frac{3\epsilon}{2n}$. Therefore by the union bound, $\Pr[\mathsf{Fail}] \le 3\epsilon$.*

Lemma 4.2 (Attack is efficient)**.** $\Pr_{\mathcal{EXEC}}[\mathsf{Long}] \le 10\epsilon$.

4.1 Success of attack: proof for Lemma 4.1

Lemma 4.1 follows from the following stronger result:

Lemma 4.3. *Let i be even and let $B = (r_B, H_B)$ be some fixing of Bob's view in an execution up to the i^{th} message sent by him, and let M, I be some fixing of the messages exchanged and query/answer pairs learned by Eve in this execution such that $\Pr_{\mathcal{EXEC}(M,I)}[\mathsf{Good}(M, I) \mid B] > 0$. Then,*

$$\Pr_{\mathcal{GEXEC}(M,I)}[\mathsf{Fail}_i \mid B] \le \frac{3\epsilon}{2n} \quad .$$

That is, the probability that Fail_i happens is at most $\frac{3\epsilon}{2n}$ conditioning on Eve's information equalling (M, I), Bob's view of the execution equalling B and $\mathsf{Good}(M, I)$.

We fist see why Lemma 4.3 implies Lemma 4.1.

Proof of Lemma 4.1 from Lemma 4.3. Lemma 4.3 implies that in particular for every even i,

$$\Pr_{\mathcal{EXEC}}[\mathsf{Fail}_i \mid \mathsf{Good}_i] \le \frac{3\epsilon}{2n} \quad ,$$

where Good_i denotes the event $\mathsf{Good}(M, I)$ where M, I are Eve's information just before the i^{th} round. But since Fail_i is the event that Eve fails at round i *for the first time*, Fail_i implies Good_i and hence $\Pr_{\mathcal{EXEC}}[\mathsf{Fail}_i] \le \Pr_{\mathcal{EXEC}}[\mathsf{Fail}_i \mid \mathsf{Good}_i]$, establishing the statement of Lemma 4.1 for every even i. By symmetry, the analog of Lemma 4.3 for odd i also holds with the roles of Alice and Bob reversed, completing the proof for all i. □

4.1.1 Proof of Lemma 4.3.

Product characterization. Lemma 4.3 would be easy if the distribution $\mathcal{GEXEC}(M, I)$ would have been a *product distribution*, with the views of Alice and Bob independent from one another. Roughly speaking this is because in this case Bob has no more information than Eve on the queries Alice made in the past, and hence also from Bob's point of view, no query is more probable than ϵ/n to have been asked by Alice. Unfortunately this is not the case. However, we can show that the distribution $\mathcal{GEXEC}(M, I)$ is equal to the distribution obtained by taking some product distribution $\mathcal{A} \times \mathcal{B}$ and conditioning it on the event $\mathsf{Good}(M, I)$. (A similar observation was made by [IR], see Lemma 6.5 there.) That is, we can show the following lemma:

Lemma 4.4 (Product characterization)**.** *For every M, I denoting Eve's information up to just before the i^{th} query, if $\Pr_{\mathcal{EXEC}(M,I)}[\mathsf{Good}(M,I)] > 0$ there exist a distribution $\mathcal{A}$ (resp. $\mathcal{B}$) over Alice's (resp. Bob's) view up to that point such that*

$$\mathcal{GEXEC}(M,I) = (\mathcal{A} \times \mathcal{B}) \mid \mathsf{Good}(M,I) \quad .$$

Proof. We will show that for every pair of Alice/Bob views (A, B) in the probability space $\mathcal{EXEC}(M,I)$ that satisfy the event $\mathsf{Good}(M,I)$, $\Pr_{\mathcal{GEXEC}(M,I)}[(A,B)] = c(M,I)\alpha_A\alpha_B$ where α_A depends only on A, α_B depends only on B and $c(M,I)$ depends only on M, I. This means that if we let $\mathcal{A}$ be the distribution such that $\Pr_{\mathcal{A}}[A]$ is proportional to α_A, and $\mathcal{B}$ be the distribution such that $\Pr_{\mathcal{B}}[B]$ is proportional to α_B, then $\mathcal{GEXEC}(M,I)$ is proportional (and hence equal to) the distribution $\mathcal{A} \times \mathcal{B} \mid \mathsf{Good}(M,I)$.

Because $(A,B) \in \mathsf{SUPP}(\mathcal{GEXEC}(A,B))$, if (A,B) happens, it makes the event $\mathsf{Good}(M,I)$ hold, and so we have

$$\Pr_{\mathcal{EXEC}(M,I)}[(A,B)] = Pr_{\mathcal{EXEC}(M,I)}[(A,B) \wedge \mathsf{Good}(M,I)] = \Pr_{\mathcal{EXEC}(M,I)}[\mathsf{Good}(M,I)] \Pr_{\mathcal{GEXEC}(M,I)}[(A,B)] \ .$$

On the other hand, by definition we have $\Pr_{\mathcal{EXEC}(M,I)}[(A,B)] = \frac{\Pr_{\mathcal{EXEC}}[(A,B,M,I) \text{ happens}]}{\Pr_{\mathcal{EXEC}}[(M,I) \text{ happens}]}$ therefore:

$$\Pr_{\mathcal{GEXEC}(M,I)}[(A,B)] = \frac{\Pr_{\mathcal{EXEC}}[(A,B,M,I) \text{ happens}]}{\Pr_{\mathcal{EXEC}}[(M,I) \text{ happens}] \Pr_{\mathcal{EXEC}(M,I)}[\mathsf{Good}(M,I)]}$$

The denominator of the righthand side is only dependent on M and I. The numerator is equal to

$$2^{-|r_A|}2^{-|r_B|}2^{-\ell|Q(A)\cup Q(B)\cup Q(I)|}.$$

The reason is that the necessary and sufficient condition that $(A = (r_A, H_A), B = (r_B, H_B), M, I)$ happens is that when we choose an execution (r'_A, r'_B, H') then $r'_A = r_A$, $r'_B = r_B$ and H is consistent on the queries in $Q(A) \cup Q(B) \cup Q(I)$ with the answers specified in H_A, H_B, I. Note that this will ensure that Alice and Bob will indeed produce the transcript M. Let $\alpha_A = 2^{-|r_A|}2^{-\ell|Q(A)\setminus Q(I)|}$ and $\beta_B = 2^{-|r_B|}2^{-\ell|Q(B)\setminus Q(I)|}$. Since $(Q(A)\setminus Q(I)) \cap (Q(B)\setminus Q(I)) = \emptyset$, the numerator is equal to $2^{-|r_A|}2^{-|r_B|}2^{-\ell|Q(A)\cup Q(B)\cup Q(I)|} = \alpha_A\beta_B 2^{-\ell|Q(I)|}$. Thus indeed $\Pr_{\mathcal{GEXEC}(M,I)}[(A,B)] = c(M,I)\alpha_A\beta_B$ where $c(M,I)$ only depends on (M,I). □

Graph characterization. This product characterization implies that we can think of $\mathcal{GEXEC}$ as a distribution over random edges of some bipartite graph G. Using some insights on the way this graph is defined, and the definition of our attacking algorithm, we will show that every vertex in G is connected to most of the vertices on the other side. We then show that this implies that Bob's chance of asking a query outside of I that was asked before by Alice is bounded by $O(\epsilon/n)$.

More precisely, fixing M, I that contain Eve's view up to just before the i^{th} round, define a bipartite graph $G = (V_L, V_R, E)$ as follows. Every node $u \in V_L$ will have a corresponding view A_u of Alice that is in the support of the distribution $\mathcal{A}$ obtained from Lemma 4.4; we let the number of nodes corresponding to a view A be proportional to $\Pr_{\mathcal{A}}[A]$, meaning that $\mathcal{A}$ corresponds to the uniform distribution over the left-side vertices V_L. Similarly, every node $v \in V_R$ will have a corresponding view of Bob B_v such that $\mathcal{B}$ corresponds to the uniform distribution over V_R. We define $Q_u = Q(A_u) \setminus Q(I)$ for $u \in V_L$ to be the set of queries *outside of* I that were asked by Alice in the view A_u, and define $Q_v = Q(B_u) \setminus Q(I)$ similarly. We put an edge in the graph between u and v (denoted by $u \sim v$) if and only if $Q_u \cap Q_v = \emptyset$. Lemma 4.4 implies that the distribution $\mathcal{GEXEC}(M,I)$ is equal to the distribution obtained by letting (u,v) be a random edge of the graph G and choosing (A_u, B_v). It turns out that this graph is *dense*:

Lemma 4.5. *Let $G = (V_L, V_R, E)$ be the graph above. Then for every $u \in V_L, d(u) \geq |V_R|(1-2\epsilon)$ and for every $v \in V_R$, $d(v) \geq |V_L|(1-2\epsilon)$ where $d(w)$ is the degree of the vertex w.*

Proof. We first show that for every $w \in V_L$, $\sum_{v\in V_R, w\not\sim v} d(v) \leq \epsilon|E|$. The reason is that the probability of vertex v being chosen when we choose a random edge is $\frac{d(v)}{|E|}$ and if $\sum_{v\in V_R, w\not\sim v} \frac{d(v)}{|E|} > \epsilon$, it means that $\Pr_{(u,v)\leftarrow_{\mathrm{R}} E}[Q_w \cap Q_v \neq \emptyset] \geq \epsilon$. Hence because $|Q_w| \leq n$, by the pigeonhole principle there exists $q \in Q_w$ such that $\Pr_{(u,v)\leftarrow_{\mathrm{R}} E}[q \in Q_v] \geq \epsilon/n$. But this is a contradiction, because then q should be in I by the definition of the attack and hence cannot be in Q_w. The same argument shows that for every $w \in V_R$, $\sum_{u \in V_L, u \not\sim w} d(u) \leq \epsilon|E|$. Thus for every vertex $w \in V_L \cup V_R$, $|E^{\not\sim}(w)| \leq \epsilon|E|$ where $E^{\not\sim}(w)$ denotes the set of edges that are not adjacent to any neighbor of w (i.e., $E^{\not\sim}(w) = \{(u,v) \in E \mid u \not\sim w \wedge w \not\sim v\}$). Now the following claim proves the lemma.

Claim 4.6. *Let $G = (V_L, V_R, E)$ be a nonempty bipartite graph such that for every vertex w, $|E^{\not\sim}(w)| \leq \epsilon|E|$ for $\epsilon \leq 1/2$, then for all $u \in V_L, d(u) \geq |V_R|(1-2\epsilon)$ and for every $v \in V_R$, $d(v) \geq |V_L|(1-2\epsilon)$.*

Proof. Let $d_L = \min\{d(u) \mid u \in V_L\}$ and $d_R = \min\{d(v) \mid v \in V_R\}$. By switching the left and right sides if necessary, we may assume without loss of generality that **(*):** $\frac{d_L}{|V_R|} \leq \frac{d_R}{|V_L|}$. Thus it suffices to prove that $1-2\epsilon \leq \frac{d_L}{|V_R|}$. Suppose $1-2\epsilon > \frac{d_L}{|V_R|}$, and let $u \in V_L$ be the vertex that $d(u) = d_L < (1-2\epsilon)|V_R|$. Because for all $v \in V_R$ we have $d(v) \leq |V_L|$, thus using **(*)** we see that $|E^{\sim}(u)| \leq d_L|V_L| \leq d_R|V_R|$ where $E^{\sim}(u) = E \setminus E^{\not\sim}(u)$. On the other hand since we assumed that $d(u) < (1-2\epsilon)|V_R|$, there are more than $2\epsilon|V_R|d_R$ edges in $E^{\not\sim}(u)$, meaning that $|E^{\sim}(u)| < |E^{\not\sim}(u)|/(2\epsilon)$. But this implies

$$|E^{\not\sim}(u)| \leq \epsilon|E| = \epsilon\left(|E^{\not\sim}(u)| + |E^{\sim}(u)|\right) < \epsilon|E^{\not\sim}(u)| + |E^{\not\sim}(u)|/2,$$

which is a contradiction for $\epsilon < 1/2$ because the graph G is nonempty. □

□

Proof of Lemma 4.3 from Lemmas 4.4 and 4.5. Let B, M, I be as in Lemma 4.3 and q be Bob's query which is fixed now. By Lemma 4.4, the distribution $\mathcal{GEXEC}(M,I)$ conditioned on getting B as Bob's view is the same as $(\mathcal{A} \times \mathcal{B})$ conditioned on $\mathsf{Good}(M,I) \wedge (\mathcal{B} = B)$. By the definition of the bipartite graph $G = (V_L, V_R, E)$ it is the same as choosing a random edge $(u,v) \leftarrow_{\mathrm{R}} E$ conditioned on $B_v = B$ and choosing (A_u, B_v). We prove Lemma 4.3 even conditioned on fixing v such that $B_v = B$. Now the distribution on Alice's view is the same as choosing $u \leftarrow_{\mathrm{R}} N(v)$ to be a random neighbor of v and choosing A_u. Let $S = \{u \in V_L \mid q \in A_u\}$. Then we have:

$$\Pr_{u \leftarrow_{\mathrm{R}} N(v)}[q \in A_u] \leq \frac{|S|}{d(v)} \leq \frac{|S|}{(1-2\epsilon)|V_L|} \leq \frac{|S||V_R|}{(1-2\epsilon)|E|} \leq \frac{\sum_{u \in S} d(u)}{(1-2\epsilon)^2|E|} \leq \frac{\epsilon}{(1-2\epsilon)^2 n} < \frac{3\epsilon}{2n}$$

The second and fourth inequalities are because of Lemma 4.5. The third one is because $|E| \leq |V_L||V_R|$. The fifth one is because of the definition of the attack which asks ϵ/n heavy queries, and the sixth one is because $\epsilon = \delta/13 < 1/13$. □

4.2 Efficiency of attack: proof of Lemma 4.2

The proof of the attack's efficiency crucially uses the fact that the attack is *successful*. Recall that Eve's criteria for "heaviness" is based on the distribution $\mathcal{GEXEC}(M,I)$ that is conditioned on her not missing any queries up to this point. However, because we have proven that the event $\mathsf{Good}(M,I)$ has significant probability, queries that are heavy under this distribution are also almost as heavy under the real distribution $\mathcal{EXEC}(M,I)$, and have probability, say, at least $\epsilon/(2n)$. Intuitively this means that, on expectation, Eve will not make more than $O(n^2/\epsilon)$ such queries. The reason is that initially Alice and Bob made at most $2n$ queries that are unknown to Eve, and each query Eve makes decreases the (nonzero) expected number of unknown queries by at least $\epsilon/(2n)$. The actual argument is more subtle, and is described below.

We say that a member of the probability space $X \in \mathcal{EXEC}$ is in the event Bad, if during the running of X, at *some* point we have $\Pr_{\mathcal{EXEC}(M,I)}[\neg\mathsf{Good}(M,I)] > 1/2$, where (M,I) are the sequence of messages and Eve's set of query/answer pairs at that moment. We define the probability space $\mathcal{EXEC}(\overline{\mathsf{Bad}})$ to be the same as $\mathcal{EXEC}$ with the difference that Eve stops asking more queries whenever at some point we get $\Pr_{\mathcal{EXEC}(M,I)}[\neg\mathsf{Good}(M,I)] > 1/2$. (The behavior of Alice and Bob is the same as before.) Note that although $\mathcal{EXEC}$ and $\mathcal{EXEC}(\overline{\mathsf{Bad}})$ are the same probability spaces, because the behavior of Eve is defined differently in them, the events Long and Fail are also defined over them is different ways.[11] However, because $\mathcal{EXEC}$

[11] In fact, the event Long in $\mathcal{EXEC}(\overline{\mathsf{Bad}})$ is a subset of Long in $\mathcal{EXEC}$, and Fail in $\mathcal{EXEC}$ is a subset of Fail in $\mathcal{EXEC}(\overline{\mathsf{Bad}})$.

and $\mathcal{EXEC}(\overline{\mathsf{Bad}})$ are identical as long as Bad does not happen we have $\Pr_{\mathcal{EXEC}}[\mathsf{Bad}] = \Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Bad}]$ and more generally for every event D whose definition might depend on Eve's behavior in the system we have $\Pr_{\mathcal{EXEC}}[\mathsf{Bad} \vee D] = \Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Bad} \vee D]$.

Thus Lemma 4.2 is obtained in two steps as follows:

Step 1. We first use the success property of the attack (i.e., $\Pr_{\mathcal{EXEC}}[\mathsf{Fail}] \leq 3\epsilon/n$) to show that $\Pr_{\mathcal{EXEC}}[\mathsf{Bad}] \leq 6\epsilon$ (which means also $\Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Bad}] = \Pr_{\mathcal{EXEC}}[\mathsf{Bad}] \leq 6\epsilon$). Note that $\neg\mathsf{Good}(M, I)$ implies that Fail_i has already happened for some i, and so $\neg\mathsf{Good}(M, I)$ implies Fail.

Step 2. We then show that in $\mathcal{EXEC}(\overline{\mathsf{Bad}})$ on average Eve will not ask more than $N = \frac{4n^2}{\epsilon}$ number of queries (see Lemma 4.8). Since Long is the event that Eve asks more than $\frac{n^2}{\epsilon^2} = \frac{N}{4\epsilon}$ queries, by Markov inequality we have $\Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Long}] \leq 4\epsilon$, and therefore we will have

$$\Pr_{\mathcal{EXEC}}[\mathsf{Long}] \leq \Pr_{\mathcal{EXEC}}[\mathsf{Long} \vee \mathsf{Bad}] = \Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Long} \vee \mathsf{Bad}] \leq \Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Long}] + \Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Bad}] \leq 10\epsilon$$

.

4.2.1 Step 1: Bounding $\Pr_{\mathcal{EXEC}}[\mathsf{Bad}]$

Note that $\neg\mathsf{Good}(M, I)$ implies that Fail_i has already happened for some i, and so $\neg\mathsf{Good}(M, I)$ implies Fail. The following lemma is implied by Lemma 6.4 in [IR], but we give a proof here for sake of completeness.

Lemma 4.7. $\Pr_{\mathcal{EXEC}}[\mathsf{Bad}] \leq 6\epsilon$.

Proof. Let's assume $\Pr_{\mathcal{EXEC}}[\mathsf{Bad}] > 6\epsilon$, and we will prove $\Pr_{\mathcal{EXEC}}[\mathsf{Fail}] > 3\epsilon$ which is a contradiction. When we run the system and the attack, instead of choosing the whole randomness (for Alice, Bob, and the oracle) at the beginning, we can choose some parts of the system first (according to their final distribution), and then choose the rest of the system from their distribution conditioned on the chosen parts. The lazy evaluation of the oracle answers is a special case of this general method. Therefore we can do as follows:

1. Run the system till an arbitrary point to get (M, I) as Eve's information about the system. We pretend that till this point we have just sampled (M, I), and the rest of the system's description is not chosen yet.
2. Choose the "true" view of Alice and Bob till this point from their distribution $\mathcal{EXEC}(M, I)$.
3. Continue running the system conditioned on the views of Alice, Bob, and Eve so far.

The moment that we sample Alice and Bob's views in the second step of the mentioned method is arbitrary, and we can choose this point to be the moment that Bad happens (if it happens at all). In other words we run the game till the moment that Bad happens: $\Pr_{\mathcal{EXEC}(M,I)}[\neg\mathsf{Good}(M, I)] > 1/2$, and then will choose Alice and Bob's computation so far, and then continue the game. But if Bad never happens we sample Alice and Bob's computation just at the end.

Since $\neg\mathsf{Good}(M, I) \subset \mathsf{Fail}$ when $\Pr_{\mathcal{EXEC}(M,I)}[\neg\mathsf{Good}(M, I)] > 1/2$ happens for the first time in our execution of the system and we choose Alice and Bob's previous computation, Fail will hold for this running of the system with probability at least $1/2$. So, if $\Pr_{\mathcal{EXEC}}[\mathsf{Bad}] > 6\epsilon$, then with probability at least $(6\epsilon)\frac{1}{2} = 3\epsilon$ the event Fail holds in the system at the end, which is not possible. ☐

4.2.2 Step 2: Bounding $\Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Long}]$

In this section the underlying sample space will always be $\mathcal{EXEC}(\overline{\mathsf{Bad}})$ even if this is not explicitly stated. Note however that Alice and Bob behave identically in $\mathcal{EXEC}$ and $\mathcal{EXEC}(\overline{\mathsf{Bad}})$. Let $\gamma = \frac{\epsilon}{2n}, N = \frac{2n}{\gamma} = \frac{4n^2}{\epsilon}$ be as defined above. The following lemma shows that $\Pr_{\mathcal{EXEC}(\overline{\mathsf{Bad}})}[\mathsf{Long}] \leq 4\epsilon$ where Long is the event of asking more than $\frac{N}{4\epsilon} = \frac{n^2}{\epsilon^2}$ number of queries.

Lemma 4.8. *The expected number of queries asked by Eve in $\mathcal{EXEC}(\overline{\mathsf{Bad}})$ is at most $N = \frac{2n}{\gamma} = \frac{4n^2}{\epsilon}$.*

Proof. By definition whenever Eve asks a query, it is ϵ/n heavy in the distribution $\mathcal{GEXEC}(M,I)$, and since we always have $\Pr_{\mathcal{EXEC}(M,I)}[\mathsf{Good}(M,I)] > \frac{1}{2}$ in $\mathcal{EXEC}(\overline{\mathsf{Bad}})$ as long as Eve has not stopped asking queries, therefore if q be Eve's query knowing (M,I) we have:

$$\Pr_{(A,B)\leftarrow_{\mathrm{R}}\mathcal{EXEC}(M,I)}[q \in Q(A)\cup Q(B)] \geq$$

$$\Pr_{\mathcal{EXEC}(M,I)}[\mathsf{Good}(M,I)] \Pr_{(A,B)\leftarrow_{\mathrm{R}}\mathcal{GEXEC}(M,I)}[q \in Q(A)\cup Q(B)] \geq \frac{\epsilon}{2n} = \gamma$$

Define the random variable Y_j to be 1 if Eve asks at least j queries and the j^{th} query that she makes was asked by Alice or Bob before. Clearly $\sum_j Y_j \leq 2n$ since Alice and Bob each make at most n queries, and hence

$$\sum_j \mathbb{E}[Y_j] = \mathbb{E}[\sum_j Y_j] \leq 2n \quad . \tag{1}$$

Claim: Let p_j be the probability that Eve asks the j^{th} query. Then $\mathbb{E}[Y_j] \geq p_j\gamma$.

Note that $\sum_j p_j$ is the expected number of queries asked by Eve, and the claim implies that $\sum_j p_j \leq \frac{1}{\gamma}\sum \mathbb{E}[Y_j] \leq \frac{2n}{\gamma}$, hence proving the lemma.

Proof of Claim: Define Y_j^q to be 1 if the j^{th} query that Eve asks is q and q was asked before by Alice or Bob. Then, $\mathbb{E}[Y_j] = \sum_q \mathbb{E}[Y_j^q]$. Let $\mathcal{O}_j$ be a random variable that whenever there is a j^{th} query asked by Eve, it denotes the information (i.e., transcript and query/answer pairs) that Eve has up to the point when it makes its j^{th} query. In an execution where Eve makes less than j queries, we define $\mathcal{O}_j = \perp$. Note that the j^{th} query of Eve is determined by $\mathcal{O}_j$ which we denote by $q(\mathcal{O}_j)$ (and we define $q(\perp) = \perp$). Let $\mathcal{W}_j = \mathsf{SUPP}(\mathcal{O}_j) \setminus \{\perp\}$, and so we will have:

$$\mathbb{E}[Y_j^q] = \sum_{\substack{L\in\mathcal{W}_j\\ q(L)=q}} \Pr[\mathcal{O}_j = L]\Pr[q \text{ asked before by Alice or Bob} \mid \mathcal{O}_j = L] \quad .$$

But by definition, if $q(L) = q$ we have $\Pr[q \text{ is asked before by Alice or Bob} \mid \mathcal{O}_j = L] \geq \gamma$. Meaning that $\mathbb{E}[Y_j^q] \geq \gamma \sum_{\substack{L\in\mathcal{W}_j\\ q=q(L)}} \Pr[\mathcal{O}_j = L]$, and hence

$$\mathbb{E}[Y_j] \geq \gamma \sum_{q\neq\perp} \sum_{\substack{L\in\mathcal{W}_j\\ q=q(L)}} \Pr[\mathcal{O}_j = L] =$$

$$\gamma \sum_{L\in\mathcal{W}_j} \Pr[\text{Eve queries some } q \text{ as its } j^{\text{th}} \text{ query} \mid \mathcal{O}_j = L]\Pr[\mathcal{O}_j = L] = \gamma p_j \quad .$$

□

5 Completing the proof

To complete the proof, we need to **(a)** show how to handle protocols that are not necessarily in normal form and **(b)** show how Eve can recover the secret once she knows the intersection queries. Task **(b)** was already achieved by [IR, Theorem 6.2] (although we give a more efficient implementation here). [IR] also showed how one can achieve task **(a)** using a general "compiler" that transforms general protocols to normal form. However that transformation has a quadratic blowup in efficiency that we cannot afford. We now sketch how our attack can be extended to handle general protocols without incurring this cost; see Section 5.1 for the full proof.

5.1 Removing the normal form assumption

In this section we show how to get an attack of the same $(\frac{13n}{\delta})^2$ complexity finding all the intersection queries of Alice and Bob for a more general form of protocols. The proof has the following two steps:

1. We extend the result with the same complexity of $(\frac{13n}{\delta})^2$ queries for the attack to the "seminormal" protocols by a bit more careful analysis of the same attack given above. A seminormal protocol is a protocol in which Alice and Bob can ask either zero or one query in each of their rounds. Again Alice and Bob ask at most n queries each, but the number of rounds R can be arbitrary larger than n.

2. Any protocol can be changed into a seminormal protocol without increasing n or loosing the security. Suppose i is a round in the original protocol in which Bob is going to ask $k \le n$ number of queries (k is not known to Eve or Alice) and then send the message m_i to Alice. In the new protocol, this round will be divided into $2n-1$ sub-rounds. In the j^{th} sub-round (of this round) if j is even, Alice will just send the message $\perp$ to Bob. So let j is an odd number. If $j \le k$, Bob will ask his j^{th} query which he was going to ask in the i^{th} round of the original protocol, and if $j > k$ he asks no query. If $j < 2n-1$, Bob sends also the message $\perp$ to Alice in the sub-round j, and if $j = 2n-1$ he sends his message m_i to Alice. It is clear that this artificial change only increases the number of rounds and will not give Eve any extra information, and therefore it is as secure as the original protocol. In the actual attack, Eve will pretend that Alice and Bob are sending the extra $\perp$ messages to each other in the sub-rounds and will attack the protocol in the seminormal form, and as we will prove she finds all the intersection queries with probability $1-\delta$ using $(\frac{13n}{\delta})^2$ number of queries.

Attack for seminormal protocols. Now we assume that Alice and Bob run a seminormal protocol in which each of them asks at most n number of oracle queries. We prove that the same attack of Section 3 finds all the intersection queries with probability $1-\delta$ using $(\frac{13n}{\delta})^2$ number of queries. We only show that the attack is successful in finding all the intersection queries and the same argument as before shows that the efficiency follows from the success property. Let BFail_i be the event that Bob's i^{th} query is the first intersection query out of I, and similarly let AFail_i be the event that Alice's i^{th} query is the first intersection query out of I.

Lemma 5.1. *For every $1 \le i \le n$, we have* $\Pr[\mathsf{BFail}_i] \le \frac{3\epsilon}{2n}$ *and* $\Pr[\mathsf{AFail}_i] \le \frac{3\epsilon}{2n}$.

Before proving the lemma note that it show $\Pr[\mathsf{Fail}] = \Pr[\bigvee_i(\mathsf{BFail}_i \vee \mathsf{AFail}_i)] \le \sum_{1\le i\le n}\Pr[\mathsf{BFail}_i] + \Pr[\mathsf{AFail}_i] \le 3\epsilon$.

Proof. Let the event BGood_i (in the seminormal protocol) be the event that when Bob asks his i^{th} query, the intersection queries of Alice and Bob are all inside I (the set of query/answer pairs known to Eve). Let $\mathcal{B}_i$ be the random variable denoting the view of Bob,till he asks his i^{th} query and let $q(B_i)$ be his i^{th} query determined by B_i. Also let $(M,I)_i$ be the random variable denoting Eve's information before Bob asks $q(B_i)$.

So we have $\Pr[\mathsf{BFail}_i] = \sum_{(B_i,M,I)\in\mathsf{SUPP}(\mathcal{B}_i\times(M,I)_i)}\Pr_{\mathcal{EXEC}}[B_i,M,I]\Pr_{\mathcal{EXEC}}[\mathsf{BFail}_i \mid B_i,M,I]$. But since BFail_i implies BGood_i we have $\Pr_{\mathcal{EXEC}}[\mathsf{BFail}_i \mid B_i,M,I] \le \Pr_{\mathcal{EXEC}}[\mathsf{BFail}_i \mid B_i,M,I,\mathsf{BGood}_i]$ and by definition we have $\Pr_{\mathcal{EXEC}}[\mathsf{BFail}_i \mid B_i,M,I,\mathsf{BGood}_i] = \Pr_{\mathcal{GEXEC}(M,I)}[\mathsf{BFail}_i \mid B_i]$. But the proof of Lemma 4.3 already showed that $\Pr_{\mathcal{GEXEC}(M,I)}[\mathsf{BFail}_i \mid B_i] \le \frac{3\epsilon}{2n}$ because it only used the fact that Alice and Bob ask at most n queries each and did not depend on the number of rounds. Hence it holds that

$$\Pr_{\mathcal{EXEC}}[\mathsf{BFail}_i] \le \sum_{(B_i,M,I)\in\mathsf{SUPP}(\mathcal{B}_i\times(M,I)_i)}\Pr_{\mathcal{EXEC}}[B_i,M,I]\frac{3\epsilon}{2n} = \frac{3\epsilon}{2n} \quad .$$

□

5.2 Finding the secret

Now, we turn to the question of finding the secret. Theorem 6.2 in [IR] shows that once one finds all the intersection queries, with $O(n^2)$ more queries they can also find the actual secret. Here we use the properties of our attack to show that we can do so even without asking more queries.

Theorem 5.2. *Assume that the total number of queries asked by Alice and Bob is at most n each, and their outputs agree with probability at least ρ having access to a random oracle. Then for every $0 < \delta < 1$ there is an adversary Eve asking at most $(\frac{16n}{\delta})^2$ number of queries such that Eve's output agrees with Bob's output with probability at least $\rho - \delta$.*

Proof. Let assume that in the last round of the protocol Alice sends a special message LAST to Bob. In order to find the secret Eve runs the attack of Section 3.1 and at the end (when Alice has sent LAST and Eve has asked her queries from the oracle), Eve samples $(A, B) \leftarrow_{\mathrm{R}} \mathcal{GEXEC}(M, I)$ (where (M, I) is Eve's information at the moment) and outputs the secret $s(A)$ determined by Alice's view A. Now we prove that her secret agrees with Bob's secret with probability at least $\rho - 16\epsilon$ and the theorem follows by setting $\delta = \epsilon/16$.

Let the random variables $\mathcal{A}, \mathcal{B}, \mathcal{E}$ be in order the view of Alice, Bob, and Eve at the end of the game. Let $\hat{\mathcal{A}}$ be the random variable generated by sampling $(A, B) \leftarrow_{\mathrm{R}} \mathcal{GEXEC}(M, I)$ where M, I are the information specified in $\mathcal{E}$ and choosing A from it. (So $s(\hat{\mathcal{A}})$ is Eve's output.) We will prove that $\mathsf{SD}((\mathcal{A}, \mathcal{B}, \mathcal{E}), (\hat{\mathcal{A}}, \mathcal{B}, \mathcal{E})) \leq 16\epsilon$ which shows that $|\Pr[s(\mathcal{A}) = s(\mathcal{B})] - \Pr[s(\hat{\mathcal{A}}) = s(\mathcal{B})]| \leq 16\epsilon$. For $(A, B, E) \in \mathsf{SUPP}(\mathcal{A} \times \mathcal{B} \times \mathcal{E})$ we say the event $\mathsf{Good}(A, B, E)$ holds if A and B do not have any intersection query out of I where $(M, I) = E$. The proof follows from the following three claims:

1. $\Pr[\neg\mathsf{Good}(\mathcal{A}, \mathcal{B}, \mathcal{E})] \leq 13\epsilon$.
2. $\Pr[\neg\mathsf{Good}(\hat{\mathcal{A}}, \mathcal{B}, \mathcal{E})] \leq \epsilon$.
3. $\mathsf{SD}((\mathcal{A}, \mathcal{B}, \mathcal{E}) \mid \mathsf{Good}(\mathcal{A}, \mathcal{B}, \mathcal{E}), (\hat{\mathcal{A}}, \mathcal{B}, \mathcal{E}) \mid \mathsf{Good}(\hat{\mathcal{A}}, \mathcal{B}, \mathcal{E})) \leq 2\epsilon$.

The first claim follows from the proof of Theorem 3.1. The second claim is true because after fixing $\mathcal{E} = (M, I)$, the random variable $\hat{\mathcal{A}}$ is independent of $\mathcal{B}$, and if we fix $\mathcal{B} = B$ any query of $Q(B)$ has chace of at most ϵ/n of being in $Q(\hat{\mathcal{A}})$ and there are at most n such queries.

So we only need to prove the third claim which we do even conditioned on fixing $\mathcal{B} = B$ and fixed $\mathcal{E} = E = (M, I)$. The claim follows from Lemma 4.5. Let $G = (V_L, V_R, D)$ be the graph characterization of $\mathcal{GEXEC}(M, I)$. Hence we have:

- The distribution of $\mathcal{A}$ in $(\mathcal{A}, B, E) \mid \mathsf{Good}(\mathcal{A}, B, E)$ is the same as choosing $v \in V_R$ such that $B_v = B$ and then choosing a random neighbor of it $u \leftarrow_{\mathrm{R}} N(v)$ and getting A_u.
- The distribution of $\hat{\mathcal{A}}$ in $(\hat{\mathcal{A}}, B, E) \mid \mathsf{Good}(\hat{\mathcal{A}}, B, E)$ is the same as choosing $v \in V_R$ such that $B_v = B$ and then choosing a random edge $(u, v') \leftarrow_{\mathrm{R}} D$ conditioned on $u \sim v$ and then getting A_u. This is the same as choosing $u \in N(v)$ at random with probability proportional to $d(u)$.

The two above distributions have statistical distance at most 2ϵ even for a fixed v such that $B_v = B$. The first distribution chooses $u \in N(v)$ uniformly at random, but the second distribution chooses $u \in N(v)$ with probabilities proportional to their degrees. But since for every $u \in V_L$, $(1 - 2\epsilon)|V_R| \leq d(u) \leq |V_R|$, (and $\epsilon < 1/10$) one can easily show that the statistical distance is bounded by 2ϵ. ☐

Acknowledgements. We thank Russell Impagliazzo for useful discussions, and also for his warning that attempting to prove an $O(n^2)$ bound for this problem leads naturally to conjecturing (and even conjecturing that you proved) intermediate results that are simply not true. He was much more prescient than we realized at the time.

References

[BMG] B. Barak and M. Mahmoody-Ghidary. Merkle Puzzles are Optimal. *Arxiv preprint arXiv:0801.3669v1*, 2008. Preliminary version of this paper. Version 1 contained a bug that is fixed in this version.

[BGP] M. Bellare, O. Goldreich, and E. Petrank. Uniform Generation of NP-Witnesses Using an NP-Oracle. *Inf. Comput*, 163(2):510–526, 2000.

[BR] M. Bellare and P. Rogaway. Random oracles are practical: A paradigm for designing efficient protocols. In *Proceedings of the First Annual Conference on Computer and Communications Security*, pages 62–73. ACM, November 1993.

[BBE] Bennett, Brassard, and Ekert. Quantum Cryptography. *SCIAM: Scientific American*, 267, 1992.

[BGI] E. Biham, Y. J. Goren, and Y. Ishai. Basing Weak Public-Key Cryptography on Strong One-Way Functions. In R. Canetti, editor, *TCC*, volume 4948 of *Lecture Notes in Computer Science*, pages 55–72. Springer, 2008.

[BS] G. Brassard and L. Salvail. Quantum Merkle Puzzles. In *International Conference on Quantum, Nano and Micro Technologies (ICQNM)*, pages 76–79. IEEE Computer Society, 2008.

[CM] Cachin and Maurer. Unconditional Security Against Memory-Bounded Adversaries. In *CRYPTO: Proceedings of Crypto*, 1997.

[CGH] R. Canetti, O. Goldreich, and S. Halevi. The Random Oracle Methodology, Revisited. In *Proc. 30th STOC*, pages 209–218. ACM, 1998.

[DH] W. Diffie and M. Hellman. New Directions in Cryptography. *IEEE Transactions on Information Theory*, IT-22(6):644–654, Nov. 1976.

[Gro] L. K. Grover. A Fast Quantum Mechanical Algorithm for Database Search. In *Annual Symposium on Theory of Computing*, pages 212–219, 22–24 May 1996.

[IR] R. Impagliazzo and S. Rudich. Limits on the provable consequences of one-way permutations. In *Proc. 21st STOC*, pages 44–61. ACM, 1989. Full version available from Russell Impagliazzo's home page.

[Mer] R. Merkle. Secure communications over insecure channels. *Communications of the ACM*, 21(4):294–299, 1978.

[RSA] R. L. Rivest, A. Shamir, and L. M. Adleman. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Communications of the ACM*, 21(2):120–126, Feb 1978.

[Sot] M. Sotakova. Breaking One-Round Key-Agreement Protocols in the Random Oracle Model. Cryptology ePrint Archive, Report 2008/053, 2008. `http://eprint.iacr.org/`.

A The error in the previous proof

The manuscript [BMG] contained a wrong proof for Theorem 1.1. The idea of that proof was to describe the attack for another setting in which Alice and Bob will get independent oracle answers for the *same* query q if they ask it before Eve. The experiment was called Imag and the original experiment (which oracle answers are the same for everyone) was called Real. The point is that if we bound the probability of (the first) missing of an intersection query in either of Imag or Real it bounds that probability in the other setting as well, because before that event the experiments are the same. It was claimed in Lemma 5.5 of [BMG] that conditioned on Eve's information till some point, the distribution of Alice and Bob's computations before that point are independent, but it was not correct. This distribution will be in fact distributed on a *product set*, but although it was originally a flat distribution (each element has the same probability), conditioning on Eve's information will introduce *reweighing* of certain elements that cause the distribution to be non-product. Below we explain this in more detail.

Let $E = (M, I)$ be Eve's information and B describes Bob's computation, and we want to sample Alice's computation A (in Imag) conditioned on (E, r_B). Lemma 5.4 of [BMG] claimed correctly that the consistency of (E, A) is necessary and sufficient for the consistency of (E, A, B) (i.e. $\mathsf{SUPP}(A \mid E) \times \mathsf{SUPP}(B \mid E) = \mathsf{SUPP}(A \times B \mid E)$), but it does not mean that r_A can be chosen *uniformly* at random, and the correct distribution actually depends on B.

One way to see why there is such dependency is to compute $\Pr[(A,B,M,I)$ happen in Imag experiment$]$ where Alice and Bob have no private intersection query: $Q(A) \cap Q(B) \subset I$. Let r_A, r_B be the length of the (original) randomness of Alice and Bob which does not describe their oracle answers (In [BMG] the parties' randomness had the oracle answers as well). Now as opposed to what we had in Lemma 4.4 (of this paper) this probability is equal to: $\Pr[(A,B,M,I)$ happen in Imag experiment$] = 2^{-|r_A|}2^{-|r_B|}2^{-\ell|Q(A)\cup Q(B)\cup I|}2^{-\ell k}$ where k is the number of intersection queries of Alice and Bob (according to (A,B,M,I)) which Eve has asked that query later at some point. The reason is that for such queries we have chosen the answer randomly at *two* points (i.e., when Alice asked it and when Bob asked it), and then when Eve asked it for the last time there was no coin tossing for getting the answer. It is different from the case for other queries which, say, Alice asked it first, Eve asked it second, and Bob asked it at last. In the latter case we only choose the random answer for Alice and no randomness is used later. The term $2^{-\ell k}$ makes the probability to be dependent on which order Alice and Bob ask their queries in the computation described by (A,B,M,I). The positive (misleading) thing about $\Pr[(A,B,M,I)$ happen in Imag experiment$]$ is that when we do not have necessarily $Q(A) \cap Q(B) \subset I$, a secret intersection query for Alice and Bob contributes $2^{-2\ell}$ to the probability which can be divided into two parts $2^{-\ell} \times 2^{-\ell}$ between Alice and Bob, but as we said for intersection queries that Eve also asks the query later the contribution of that query to the probability depends on the order that parties ask it.

Therefore, when we want to know the distribution of Alice in Imag conditioned on (M,I), if q is asked in Alice's view A and also we have $q \in I$, then A is less probable in the case Bob has asked q before Eve compared to the case that it is asked after Eve.