Secret-Extraction Attacks against Obfuscated Instantaneous Quantum Polynomial-Time Circuits

David Gross[*]
Institute for Theoretical Physics, University of Cologne, D-50937 Köln, Germany

Dominik Hangleiter[†]
Simons Institute for the Theory of Computing, University of California at Berkeley, Berkeley, California 94720, USA
and Joint Center for Quantum Information and Computer Science (QuICS), University of Maryland and National Institute of Standards and Technology (NIST), College Park, Maryland 20742, USA

(Received 15 April 2024; revised 22 October 2024; accepted 13 February 2025; published 22 April 2025)

Quantum computing devices can now perform sampling tasks that, according to complexity-theoretic and numerical evidence, are beyond the reach of classical computers. This raises the question of how one can efficiently verify that a quantum computer operating in this regime works as intended. In 2008, Shepherd and Bremner proposed a protocol in which a verifier constructs a unitary from the comparatively easy-to-implement family of so-called *IQP circuits* and challenges a prover to execute it on a quantum computer. The challenge problem is designed to contain an obfuscated secret, which can be turned into a statistical test that accepts samples from a correct quantum implementation. It was conjectured that extracting the secret from the challenge problem is NP-hard, so that the ability to pass the test constitutes strong evidence that the prover possesses a quantum device and that it works as claimed. Unfortunately, about a decade later, Kahanamoku-Meyer found an efficient classical secret-extraction attack. Bremner, Cheng, and Ji very recently followed up by constructing a wide-ranging generalization of the original protocol. Their *IQP Stabilizer Scheme* has been explicitly designed to circumvent the known weakness. They also suggested that the original construction can be made secure by adjusting the problem parameters. In this work, we develop a number of secret-extraction attacks that are effective against both new approaches in a wide range of problem parameters. In particular, we find multiple ways to recover the 300-bit secret hidden in a challenge data set published by Bremner, Cheng, and Ji. The important problem of finding an efficient and reliable verification protocol for sampling-based proofs of quantum advantage thus remains open.

DOI: 10.1103/PRXQuantum.6.020314

I. INTRODUCTION

A central challenge in the field of quantum advantage is to devise efficient quantum protocols that are both classically intractable and classically verifiable, while minimizing the experimental effort required for an implementation. The paradigmatic approach satisfying these first conditions is to solve public key cryptography schemes using Shor's algorithm. However, the quantum resources required in the cryptographically secure regime are enormous, using thousands of qubits and millions of gates (see, e.g., Refs. [1,2]). Reducing the required resources, interactive proofs of computational quantumness have been proposed, which make use of classically or quantum secure cryptographic primitives [3–5]. Again, however, their implementation requires arithmetic operations, putting the advantage regime far beyond the reach of current technology [6].

A different approach to demonstrations of quantum advantage has focused on simple protocols based on sampling from the output of random quantum circuits [7–12]. These require a significantly smaller amount of qubits and gates, and seem to be classically intractable even in the presence of noise on existing hardware [13–16]. However, they are not efficiently verifiable (for a discussion, see Ref. [17]) and present-day experiments are already outside of the regime in which the samples can be efficiently checked.

The key property that makes random quantum sampling so much more feasible compared to cryptography-based

[*] Contact author: david.gross@thp.uni-koeln.de
[†] Contact author: mail@dhangleiter.eu

approaches is their apparent lack of structure in the sampled distribution. At the same time, this is also what seems to thwart classical verifiability. Yamakawa and Zhandry [18] have made significant progress by showing that there are also highly unstructured NP-problems based on random oracles, which can be efficiently solved by a quantum computer and checked by a classical verifier. Conversely, one may wonder whether it is possible to introduce just enough structure into random quantum circuits to make their classical outputs efficiently verifiable while keeping the resource requirements low [19–21]. An early and influential idea of this type dating back to 2008 is that of Shepherd and Bremner [19].

Shepherd and Bremner proposed a sampling-based scheme based on so-called *instantaneous quantum polynomial-time (IQP)* circuits. In the IQP paradigm, one can only execute gates that are diagonal in the X basis. They designed a family of IQP circuits based on *quadratic residue codes (QRCs)* the output distribution of which has high weight on bit strings $\mathbf{x} \in \mathbb{F}_2^n$ that are contained in a hyperplane $\mathbf{s}^T\mathbf{x} = 0 \mod 2$. The normal vector $\mathbf{s} \in \mathbb{F}_2^n$ may be chosen freely but its value is not apparent from the circuit description. In this way, a verifier can design a circuit that hides a *secret* $\mathbf{s}$. The verifier then challenges a prover to produce samples such that a significant fraction of them lie in the hyperplane orthogonal to $\mathbf{s}$. At the time, the only known way to efficiently meet the challenge was for the prover to collect the samples by implementing the circuit on a quantum computer. More precisely, Shepherd and Bremner conjectured that it was an NP-hard problem to recover the secret from the circuit description. They challenged the community to generate samples that have high overlap with a secret 244-bit string—corresponding to a 244-qubit experiment [22].

Unfortunately, in 2019, Kahanamoku-Meyer [23] solved the challenge and recovered the secret string. The paper provided evidence that the attack has only quadratic running time for the QRC construction.

Recently, Bremner, Cheng, and Ji [24] have made new progress on this important problem. They propose a wide-ranging generalization of the construction—the *IQP Stabilizer Scheme*—which circumvents Kahanamoku-Meyer's analysis. They also conjecture that an associated computational problem—the *Hidden Structured Code (HSC) problem*—cannot be solved efficiently classically for some parameter choices and pose a challenge for an IQP experiment on 300 qubits, corresponding to a 300-bit secret. Finally, they also extend the QRC-based construction to parameter regimes in which Kahanamoku-Meyer's ansatz fails.

Here, we show that the scheme is still vulnerable to classical cryptanalysis by devising a number of secret-extraction attacks against obfuscated IQP circuits. Our first approach, the *Radical Attack* instantly recovers the 300-bit secret of the challenge from the circuit description. We analyze the Radical Attack in detail and give conditions under which we expect the ansatz to work. The theory is tested on 100 000 examples generated by a software package provided as part of Ref. [24] and is found to match the empirical data well. We also observe that for the Extended QRC construction, the Radical Attack and the approach of Kahanamoku-Meyer complement each other almost perfectly, in the sense that for every parameter choice, exactly one of the two works with near certainty.

In Sec. V, we sketch a collection of further approaches for recovering secrets hidden in IQP circuits. Concretely, we propose two extensions of Kahanamoku-Meyer's idea, which we call the *Lazy Linearity Attack* and the *Double Meyer*. The Double Meyer Attack is effective against the Extended QRC construction for all parameter choices and we expect that its running time is at most quasipolynomial on all instances of the IQP Stabilizer Scheme. Finally, we introduce *Hamming's Razor*, which can be used to identify redundant rows and columns of the matrix that were added as part of the obfuscation procedure. For the challenge data set, this allows us to recover the secret in an alternative fashion and we expect it to reduce the load on further attacks in general.

The important problem of finding cryptographic obfuscation schemes for the efficient classical verification of quantum circuit implementations therefore remains open.

We begin by setting up some notation in Sec. II, then recall the IQP Stabilizer Scheme in Sec. III, describe and analyze the Radical Attack in Sec. IV, and sketch further exploits in Sec. V.

II. NOTATION AND DEFINITIONS

We mostly follow the notation of Ref. [24]. This means using boldface for matrices $\mathbf{M}$ and column vectors $\mathbf{v}$ (though basis-independent elements of abstract vector spaces are set in lightface). We write $\mathbf{M}_i$ for the ith column of a matrix and $\mathbf{v}_i$ for the ith coefficient of a column vector. By the *support of a set* $S \subset \mathbb{F}_2^m$, we mean the union of the supports of its elements:

$$\operatorname{supp} V = \{i \in [1, m] \mid \exists \mathbf{v} \in V, \quad \mathbf{v}_i \neq 0\}.$$

We use $\mathbf{e}^i$ for the standard basis vector $(\mathbf{e}^i)_j = \delta_{ij}$. The all-ones vector is $\mathbf{1}$ and for a set S, $\mathbf{1}_S$ is the "indicator function on S," i.e., the ith coefficient of $\mathbf{1}_S$ is 1 if $i \in S$ and 0 else. The *Hamming weight* of a vector $\mathbf{v} \in \mathbb{F}_2^n$ is $|\mathbf{v}| := \sum_{i\in[n]} \mathbf{v}_i$.

A. Symmetric bilinear forms

Compared to Ref. [24], we use slightly more geometric language. The relevant notions from the theory of symmetric bilinear forms, all standard, are briefly recapitulated here.

Let V be a finite-dimensional vector space over a field $\mathbb{F}$, endowed with a symmetric bilinear form $\beta(\cdot,\cdot)$. The *orthogonal complement* of a subset $W \subset V$ is

$$W^\perp = \{x \in V \mid \beta(x, w) = 0 \quad \forall w \in W\}.$$

The *radical* of V is $\mathrm{rad}\, V = V \cap V^\perp$, which is the space comprising elements $x \in V$ such that the linear function $\beta(x, \cdot)$ vanishes identically on V. The space V is *nondegenerate* if $\mathrm{rad}\, V = \{0\}$. In this case, for every subspace $W \subset V$, we have that $\dim W^\perp + \dim W = \dim V$. The subspace W is *isotropic* if $W \subset W^\perp$. The above dimension formula implies that isotropic subspaces W of a nondegenerate space V satisfy $\dim W \leq \frac{1}{2} \dim V$.

Let $\{b^{(i)}\}_{i=1}^k$ be a basis of V. Expanding vectors $x, y \in V$ in the basis,

$$\beta(x,y) = \beta\left(\sum_i \mathbf{c}_i b^{(i)}, \sum_j \mathbf{d}_j b^{(j)},\right)$$
$$= \sum_{ij} \mathbf{c}_i \mathbf{d}_j \, \beta\left(b^{(i)}, b^{(j)}\right) = \mathbf{c}^T \mathbf{M} \mathbf{d},$$

where $\mathbf{c}, \mathbf{d} \in \mathbb{F}^k$ are column vectors containing the expansion coefficients of x and y, respectively, and the *matrix representation* $\mathbf{M}$ of β has elements

$$\mathbf{M}_{ij} = \beta\left(b^{(i)}, b^{(j)}\right).$$

A vector $x \in V$ lies in the radical if and only if its coefficients $\mathbf{c}$ lie in the kernel of $\mathbf{M}$.

Now assume that $V \subset \mathbb{F}^m$. The *standard form* on $\mathbb{F}^m$ is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^m \mathbf{x}_i \mathbf{y}_i.$$

The standard form is *nondegenerate* on $\mathbb{F}^m$ but will in general be degenerate on subspaces. Let $\mathbf{H}$ be an $m \times k$ matrix and let $W = \mathrm{range}\, \mathbf{H} \subset V$ be its column span. The restriction of the standard form to W can then be "pulled back" to $\mathbb{F}^k$ by mapping $\mathbf{c}, \mathbf{d} \in \mathbb{F}^k$ to

$$\langle \mathbf{Hc}, \mathbf{Hd} \rangle = (\mathbf{Hc})^T (\mathbf{Hd}) = \mathbf{c}^T \left(\mathbf{H}^T \mathbf{H}\right) \mathbf{d} = \mathbf{c}^T \mathbf{G} \mathbf{d},$$

where $\mathbf{G} = \mathbf{H}^T \mathbf{H}$ is the *Gram matrix* associated with $\mathbf{H}$. We frequently use the fact that in this context,

$$\mathbf{d} \in \ker \mathbf{G} \quad \Leftrightarrow \quad \mathbf{Hd} \in \mathrm{rad}\, W. \tag{1}$$

III. THE IQP STABILIZER SCHEME

A. Hiding a secret string in an IQP circuit

The IQP Stabilizer Scheme of Shepherd and Bremner, described here following the presentation in Ref. [24], uses the tableau representation of a collection of m Pauli matrices on n qubits as an $m \times 2n$ binary matrix. Since IQP circuits are diagonal in the X basis, we restrict to X-type Pauli matrices, which are described by $m \times n$ matrices with elements in $\mathbb{F}_2$. The tableau matrix $\mathbf{H} \in \mathbb{F}_2^{m \times n}$ determines an IQP Hamiltonian $H = \sum_{i \in [m]} (\prod_{j \in [n]} X_j^{\mathbf{H}_{ij}})$, with associated IQP circuit ω^H defined in terms of a phase ω. Choosing $\omega = e^{i\pi/4}$, Shepherd and Bremner observe that the full stabilizer tableau of the state $\omega^H |0\rangle$ can be expressed in terms of $\mathbf{H}$ and use this fact to find IQP circuits the output distributions of which have high weight on a subspace $S_\mathbf{s} = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{s} \rangle = 0\}$ determined by a secret string $\mathbf{s}$.

This is ingeniously achieved as follows. For $\mathbf{s} \in \mathbb{F}_2^n$, obtain $\mathbf{H_s}$ from $\mathbf{H}$ by multiplying its ith row with $(\mathbf{Hs})_i$. Let $\mathcal{C}_\mathbf{s} := \mathrm{range}\, \mathbf{H_s}$. Fix some $g_{\max} \in \mathbb{N}$. A vector $\mathbf{s} \in \mathbb{F}_2^n$ is called a *secret* of $\mathbf{H}$ if

(1) the co-dimension of the radical $g := \dim \mathcal{C}_\mathbf{s} - \dim \mathrm{rad}\, \mathcal{C}_\mathbf{s} \leq g_{\max}$ and

(2) the radical is doubly even, i.e., for all $\mathbf{x} \in \mathrm{rad}\, \mathcal{C}_\mathbf{s}$, $|\mathbf{x}| = 0 \mod 4$

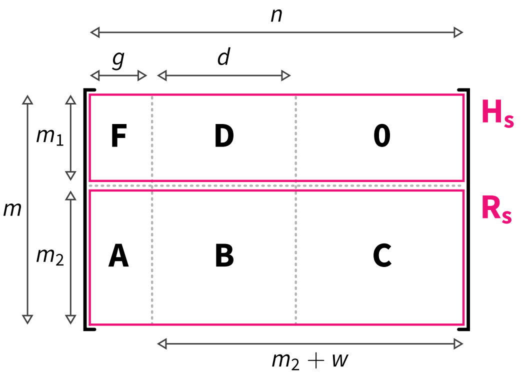
Given an IQP tableau $\mathbf{H}$ with secret $\mathbf{s}$, Shepherd and Bremner show that

$$\Pr_{\mathbf{x} \leftarrow D_\mathbf{H}} [\langle \mathbf{x}, \mathbf{s} \rangle = 0] = \frac{1}{2}(2^{-g/2} + 1), \tag{2}$$

where $D_\mathbf{H}(\mathbf{x}) = |\langle x | \omega^H | 0 \rangle|^2$ is the output distribution of ω^H. A classical verifier can then efficiently identify samples from the correct distribution by computing their mean overlap with the secret string $\mathbf{s}$.

B. Stabilizer construction

We briefly recap the specifics of how Bremner, Cheng, and Ji [24] construct a pair $(\mathbf{H}, \mathbf{s})$, comprising a generator matrix $\mathbf{H}$ and a corresponding secret $\mathbf{s}$. Before obfuscation, the matrix $\mathbf{H}$ is of the following form:



(3)

Essentially, the blocks are chosen uniformly at random, subject to the following constraints:

(1) $\mathbf{D}$ is an $m_1 \times d$ matrix. Its range is a d-dimensional doubly even isotropic subspace of $\mathbb{F}_2^{m_1}$. These

constraints are equivalent to

$$|\mathbf{D}_i| = 0 \mod 4 \quad (i = 1, \dots, d), \quad \mathbf{D}^T\mathbf{D} = 0, \\ \operatorname{rank}\mathbf{D} = d. \tag{4}$$

(2) $\mathbf{F}$ is an $m_1 \times g$ matrix. It generates a g-dimensional space that is nondegenerate and orthogonal to range $\mathbf{D}$ with respect to the standard form on $\mathbb{F}_2^{m_1}$. These constraints are equivalent to

$$\operatorname{rank}\mathbf{F}^T\mathbf{F} = g, \quad \mathbf{D}^T\mathbf{F} = 0. \tag{5}$$

(3) There exists a *secret* $\mathbf{s} \in \mathbb{F}_2^n$ such that the inner product between $\mathbf{s}$ and the rows of $\mathbf{H}$ is nonzero exactly for the first m_1 rows:

$$\mathbf{H}\mathbf{s} = \mathbf{1}_{[1,\dots,m_1]}. \tag{6}$$

(4) Finally, $\mathbf{R_s} = (\mathbf{A} \mid \mathbf{B} \mid \mathbf{C})$ are "redundant rows," chosen such that $\mathbf{R_s}\mathbf{s} = \mathbf{0}$ and

$$\operatorname{rank}\mathbf{H} = n. \tag{7}$$

Further comments:

(a) Introducing notation not used in Ref. [24], we split $\mathbf{R_s}$ into submatrices $\mathbf{R_s} = (\mathbf{A} \mid \mathbf{B} \mid \mathbf{C})$ according to $\mathbb{F}_2^n = \mathbb{F}_2^g \oplus \mathbb{F}_2^d \oplus \mathbb{F}_2^{n-g-d}$.

(b) It will turn out that the parameter

$$w := n - g - m_2$$

plays a central role for the performance of the Radical Attack. It may be described as measuring the degree to which the matrix $(\mathbf{B} \mid \mathbf{C})$ is "wide" rather than "tall."

(c) The range of $\mathbf{H_s}$ is the code space $\mathcal{C}_\mathbf{s}$. The range of $\mathbf{D}$ is its radical $\operatorname{rad}\mathcal{C}_\mathbf{s} \subset \mathcal{C}_\mathbf{s}$. The range of $\mathbf{F}$ is a subspace that is complementary to the radical within the code space.

(d) The rank constraint in Eq. (7) implies that

$$\operatorname{rank}\mathbf{C} = n - g - d. \tag{8}$$

(e) There are some subtleties connected to the way in which the redundant rows are generated according to Ref. [24] and the various versions of the software implementation provided: for more details, see the Appendix.

The parameters n (number of qubits), m (terms in the Hamiltonian), and g (log of the power of the statistical test) are supplied by the user, while m_1 and d are chosen randomly. The precise way in which m_1 and d are to be generated does not seem to be specified in the paper, so we take guidance from the reference implementation provided in Ref. [25]. Their `sample_parameters()` function (found in `lib/construction.py`) fixes these numbers in a two-step procedure. First, preliminary values of $m_1/2$ and d are sampled according to binomial distributions, with parameters roughly given as

$$m_1/2 \sim \operatorname{Bin}\left(N \approx \frac{m-g}{2}, p = 0.3\right), \\ d \sim \operatorname{Bin}\left(N = \left\lfloor \frac{m_1 - g}{2} \right\rfloor, p = 0.75\right). \tag{9}$$

The values are accepted if they satisfy the constraints

$$n - g \geq d \geq w. \tag{10}$$

We are not aware of a simple description of the distribution conditioned on the values passing the test. Empirically, we find that for the challenge parameters

$$n = 300, \quad m = 360, \quad g = 4, \tag{11}$$

the following values are attained most frequently:

$$m_1 = 102, \quad d = 38 \quad \Rightarrow \quad m_2 = 258, \quad w = 38. \tag{12}$$

Given $(\mathbf{H}, \mathbf{s})$, obfuscation is then performed as $\mathbf{H} \leftarrow \mathbf{PHQ}$, $\mathbf{s} \leftarrow \mathbf{Q}^{-1}\mathbf{s}$ using a random invertible matrix $\mathbf{Q}$ and a random (row) permutation $\mathbf{P}$.

Bremner, Cheng, and Ji pose the following conjecture.

Conjecture 1 (Hidden Structured Code (HSC) problem [24]). For certain appropriate choices of n, m, and g, there exists an efficiently samplable distribution over instances $(\mathbf{H}, \mathbf{s})$ from the family $\mathcal{H}_{n,m,g}$, so that no polynomial-time classical algorithm can find the secret s given n, m, and $\mathbf{H}$ as input, with high probability over the distribution on $\mathcal{H}_{n,m,g}$.

C. Extended QRC construction

In the original QRC construction of Shepherd and Bremner [19], $(\mathbf{F} \mid \mathbf{D})$ is chosen as a $q \times (q+1)/2$ QRC with prime q such that $q + 1 \mod 8 = 0$. Then, the all-ones vector $\mathbf{1}_q$, which is guaranteed to be a code word of a QRC, is appended as the first column of $(\mathbf{F} \mid \mathbf{D})$. Next, rows are added that are uniformly random, except for the first entry, which is 0. This ensures that there is a secret $\mathbf{s} = \mathbf{e}_1$. The resulting matrix is then obfuscated as above.

In the Extended QRC construction, additional redundant columns are added, essentially amounting to a nontrivial choice of $\mathbf{C}$, in order to render the algorithm of Kahanamoku-Meyer ineffective. Letting $r = (q+1)/2$, Bremner, Cheng, and Ji propose to add q redundant rows to achieve $m = 2q$ and add a redundant $\mathbf{C}$ block to achieve a width n satisfying $r \leq n \leq q + r$.

ALGORITHM 1. Radical Attack.

```
function RadicalAttack($\mathbf{H}$)
    $\mathbf{G} \leftarrow \mathbf{H}^T\mathbf{H}$
    $\mathbf{K} \leftarrow$ a column-generating matrix for $\ker \mathbf{G}$
    $S \leftarrow$ the support of the columns in $\mathbf{HK}$
    Solve the $\mathbb{F}_2$-linear system $\mathbf{Hs} = \mathbf{1}_S$
    return s
end function
```

IV. THE RADICAL ATTACK

The starting point of the attack was the empirical observation that $\mathbf{H}^T\mathbf{H}$ for the challenge-generator matrix has a much larger kernel ($\dim\ker \mathbf{H}^T\mathbf{H} = 34$) than would be expected for a random matrix of the same shape (about 1). This observation gave rise to the Radical Attack, summarized in Algorithm 1.

We have tested this ansatz against instances created by the software package provided by Bremner, Cheng, and Ji. For the parameters $(n, m, g) = (300, 360, 4)$, used for the challenge data set, the secret is recovered with probability about 99.85%. The challenge secret itself can be found using a mildly strengthened version.

We will analyze this behavior theoretically in Sec. IV A, report on the numerical findings in Sec. IV B, and, in Sec. IV C, explain why the challenge Hamiltonian requires a modified approach.

A. Performance of the attack

The analysis combines three ingredients:

(1) We will show that, with high probability, $\mathbf{H}(\ker \mathbf{G})$ is a subspace of the radical $\operatorname{rad}\mathcal{C}_\mathbf{s}$. Because $\mathbf{H}$ and thus $\mathbf{G} = \mathbf{H}^T\mathbf{H}$ are known, this means that we can access elements of the radical in a computationally efficient way.

(2) We will then show that the intersection of $\mathbf{H}(\ker \mathbf{G})$ with the radical is expected to be relatively large.

These two statements follow as Corollary 1 from a structure-preserving normal form for obfuscated generator matrices of the form given in Eq. (3), described in Lemma 1.

(3) In Lemma 2, we argue that one can expect that the nonzero coordinates that show up in this subspace coincide with the obfuscated coordinates $1 \dots m_1$.

1. A normal form for generator matrices

Recall the notion of *elementary column operations* on a matrix, as used in the context of Gaussian elimination. Over $\mathbb{F}_2$, these are (1) exchanging two columns and (2) adding one column to another one. Performing a sequence of column operations is equivalent to applying an invertible matrix from the right. We will map the generator matrix $\mathbf{H}$ to a normal form using a restricted set of column operations. These column operations preserve the properties of the blocks of $\mathbf{H}$ described in Sec. III.

To introduce the normal form, split $\mathbf{H}$ into blocks as

$$\mathbf{H} = (\hat{\mathbf{A}} \mid \hat{\mathbf{B}} \mid \hat{\mathbf{C}}), \quad \hat{\mathbf{A}} := \begin{pmatrix} \mathbf{F} \\ \mathbf{A} \end{pmatrix}, \quad \hat{\mathbf{B}} := \begin{pmatrix} \mathbf{D} \\ \mathbf{B} \end{pmatrix},$$
$$\hat{\mathbf{C}} := \begin{pmatrix} \mathbf{0} \\ \mathbf{C} \end{pmatrix}.$$

We say that an elementary column operation *is directed to the left* if it

(a) adds a columns of $\hat{\mathbf{C}}$ to another column in $(\hat{\mathbf{A}} \mid \hat{\mathbf{B}} \mid \hat{\mathbf{C}})$
(b) adds a column of $\hat{\mathbf{B}}$ to another column in $(\hat{\mathbf{A}} \mid \hat{\mathbf{B}})$
(c) adds a column of $\hat{\mathbf{A}}$ to another column in $\hat{\mathbf{A}}$, or
(d) permutes two columns within one block

The first part of the following lemma lists properties that are preserved under such column operations. The second part describes two essential simplifications to $\mathbf{H}$ that can still be achieved.

Lemma 1 (Normal form). Assume $\mathbf{H}'$ results from $\mathbf{H}$ by a sequence of column operations that are directed to the left. Then:

(1) If $\mathbf{H}$ is a block matrix of the form given in Eq. (3) and fulfills the conditions in Eqs. (5)–(7), then the same is true for $\mathbf{H}'$.
(2) It holds that:

$$\operatorname{range}(\mathbf{F}' \mid \mathbf{D}') = \operatorname{range}(\mathbf{F} \mid \mathbf{D}),$$
$$\operatorname{range}\mathbf{D}' = \operatorname{range}\mathbf{D},$$
$$\operatorname{range}(\mathbf{B}' \mid \mathbf{C}') = \operatorname{range}(\mathbf{B} \mid \mathbf{C}).$$

There is a sequence of column operations directed to the left such that:

(3) If $\operatorname{range}(\mathbf{B} \mid \mathbf{C}) = \mathbb{F}_2^{m_2}$, then $\mathbf{A}' = 0$.
(4) In any case, $\dim\ker \mathbf{B}' \geq w$.

Proof. Claims (1) and (2) follow directly from the definitions. Least trivial is the statement that the condition in Eq. (5) is preserved, so we make this one explicit. Consider the case in which the ith column of $\hat{\mathbf{B}}$ is added to the

jth column of $\hat{\mathbf{A}}$. This will change $\mathbf{F} \mapsto \mathbf{F}' = \mathbf{F} + \mathbf{D}_i\,(\mathbf{e}^j)^T$. But then, by Eq. (4),

$$\begin{aligned}(\mathbf{F}')^T\mathbf{F}' &= \mathbf{F}^T\mathbf{F} + \mathbf{e}^j\,\mathbf{D}_i^T\mathbf{F} + \mathbf{F}^T\mathbf{D}_i\,(\mathbf{e}^j)^T + \mathbf{e}^j\,\mathbf{D}_i^T\mathbf{D}_i\,(\mathbf{e}^j)^T \\ &= \mathbf{F}^T\mathbf{F}.\end{aligned}$$

Claim (3) is now immediate. Assuming the range condition, every column of $\mathbf{A}$ can be expressed as a linear combination of columns in $(\mathbf{B} \mid \mathbf{C})$. Therefore, $\mathbf{A}$ may be eliminated by column operations directed to the left.

To prove claim (4), choose a basis $\{\mathbf{b}^i\}_{i=1}^k$ for range $\mathbf{B} \cap$ range $\mathbf{C}$. Using column operations within $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$, respectively, we can achieve that the first k columns of $\mathbf{B}$ and of $\mathbf{C}$ are equal to $\mathbf{b}^1, \ldots, \mathbf{b}^k$. The first k columns of $\mathbf{B}$ can then be set to zero by subtracting the corresponding columns of $\mathbf{C}$. Using Eq. (8) and the trivial bound on $\dim(\operatorname{range}\mathbf{B} \cap \operatorname{range}\mathbf{C})$,

$$\begin{aligned}k = \dim(\operatorname{range}\mathbf{B} \cap \operatorname{range}\mathbf{C}) &\geq \operatorname{rank}\mathbf{B} + \operatorname{rank}\mathbf{C} - m_2 \\ &= (\operatorname{rank}\mathbf{B} - d) + n - g - m_2,\end{aligned}$$

so that the kernel of the resulting matrix $\mathbf{B}'$ satisfies

$$\dim\ker\mathbf{B}' = d - \operatorname{rank}\mathbf{B} + k \geq n - g - m_2 = w.$$

■

2. Accessing elements from the radical

As alluded to at the beginning of this section, the normal form implies that $\mathbf{H}(\ker\mathbf{G})$ is expected to be a subspace of $\operatorname{rad}\mathcal{C}_\mathbf{s}$, which is fairly large. More precisely, we have the following.

Corollary 1. We have that:

(1) If $\operatorname{range}(\mathbf{B} \mid \mathbf{C}) = \mathbb{F}_2^{m_2}$, then $\mathbf{H}(\ker\mathbf{G}) \subset \big((\operatorname{rad}\mathcal{C}_\mathbf{s}) \oplus 0\big)$.

(2) In any case, $\dim(\operatorname{rad}\mathcal{C}_\mathbf{s} \cap \mathbf{H}(\ker\mathbf{G})) \geq w$.

Proof. By Lemma 1 and Eq. (1), the advertised statements are invariant under column operations directed to the left.

If $\operatorname{range}(\mathbf{B} \mid \mathbf{C}) = \mathbb{F}_2^{m_2}$, we may thus assume that $\mathbf{A} = 0$, which gives

$$\mathbf{G} = \begin{pmatrix} \mathbf{F}^T\mathbf{F} & 0 & 0 \\ 0 & \mathbf{B}^T\mathbf{B} & \mathbf{B}^T\mathbf{C} \\ 0 & \mathbf{C}^T\mathbf{B} & \mathbf{C}^T\mathbf{C} \end{pmatrix}.$$

Because $\mathbf{F}^T\mathbf{F}$ has full rank, $\ker\mathbf{G} = 0 \oplus (\ker(\mathbf{B} \mid \mathbf{C}))$. But elements of this space are mapped into $\operatorname{rad}\mathcal{C}_\mathbf{s}$ under $\mathbf{H}$. This proves the first claim.

From the block form given in Eq. (3) and the fact that $\mathbf{D}$ is nondegenerate, it follows that $\mathbf{H}$ embeds $0 \oplus (\ker\mathbf{B}) \oplus 0 \subset \ker\mathbf{G}$ into $\operatorname{rad}\mathcal{C}_\mathbf{s}$. Claim (2) then follows from $\dim\ker\mathbf{B} \geq n - g - m_2$, which we may assume since the claim is invariant under column operations directed to the left. ■

If we model $\mathbf{B}$ and $\mathbf{C}$ as random matrices with elements drawn uniformly from $\mathbb{F}_2$, the probability that $\operatorname{range}(\mathbf{B} \mid \mathbf{C}) = \mathbb{F}_2^{m_2}$ can be estimated from the well-studied theory of random binary matrices. Indeed, in the limit $k \to \infty$, the probability $\rho(w)$ that a random binary $k \times (k + w)$ matrix has rank less than k is given by

$$\rho(w) = 1 - \prod_{i=w+1}^{\infty} \left(1 - 2^{-i}\right)$$

(cf. Ref. [26, Thm 3.2.1]). This expression satisfies

$$2^{-w} \leq \rho(w) \leq 2^{-w+1}, \quad \lim_{w\to\infty} \rho(w)2^w = 1 \tag{13}$$

and one may verify on a computer that 2^{-w} is an excellent multiplicative approximation to $\rho(w)$ already for $w \approx 7$. Thus, interestingly, the value of w governs the behavior of both parts of Corollary 1.

3. Reconstructing the support from random samples

We proceed to the third ingredient of the analysis—asking whether the support of the numerically obtained elements of the radical is likely to be equal to the obfuscated first m_1 coordinates.

Lemma 2. Let V be a subspace of $\mathbb{F}_2^{m_1}$. Take k elements $\{\mathbf{v}^i\}_{i=1}^k$ from V uniformly at random. The probability that $\operatorname{supp}(\{\mathbf{v}^i\}_{i=1}^k) \neq \operatorname{supp} V$ is no larger than $m_1 2^{-k}$.

Proof. Let $j \in \operatorname{supp} V$. We can find a basis $\mathbf{b}^j$ of V such that exactly $\mathbf{b}^1$ is nonzero on the jth coordinate. Therefore, for each j, exactly half the elements of V are nonzero on j. Thus, the probability that j is not contained in the support of the vectors is 2^{-k}. The claim follows from the union bound. ■

To apply the lemma to the situation at hand, let us again adopt a simple model in which the blocks of $\mathbf{H}$ are represented by uniformly random matrices. Under this model, we expect $\operatorname{supp}\operatorname{range}\mathbf{D} = [m_1]$ to hold if $d > \log_2 m_1$ and, in turn, $\operatorname{supp}\mathbf{H}(\ker\mathbf{G}) = \operatorname{supp}\operatorname{range}\mathbf{D}$ if $w > \log_2 m_1$. Again, the probability of failure decreases exponentially in the amount by which these bounds are exceeded. For the reference values in Eq. (12), $\log_2 m_1 \approx 6.67$.

While it is highly plausible that the uniform random model accurately captures the distribution of $\mathbf{H}(\ker\mathbf{G})$, this is less obvious for $\mathbf{D}$, which is constrained to have doubly even, orthogonal, and linearly independent columns. Nonetheless, it will turn out that the predictions made based on this model fit the empirical findings very well. This suggests that in the choice of $\mathbf{D}$, full support on $[m_1]$ is attained at least as fast as suggested by the random model. We leave finding a theoretical justification for this behavior as an open question [27].

The factor m_1 in the probability estimate of Lemma 2 comes from a union bound and is tight only in the unrealistic case in which for every possible choice of $\{\mathbf{v}^i\}_{i=1}^k$, at most one element is missing from the support. A less rigorous, but plausibly more realistic, estimate of the error probability is obtained if we assume that the coefficients $\mathbf{v}_j$ of random elements $\mathbf{v}$ of V are distributed independently. In this case, the error probability is $(1-2^{-k})^{m_1}$ rather than $m_1 2^{-k}$.

4. Global analysis

Combining the various ingredients, we can estimate the probability of recovering the secret given w. If all conditions are modeled independently (rather than using more conservative union bounds), the result is

$$\mathrm{Prob}[\text{success} \mid w] \approx (1-\rho(w))(1-2^{-w})^{m_1}(1-2^{-d})^{m_1}. \tag{14}$$

A number of simplifying approximations are possible. The approximate validity of some of these steps, such as dropping the "+1" in the exponent in the following displayed equation, are best verified by graphing the respective curves on a computer.

From Eq. (13),

$$\begin{aligned}\mathrm{Prob}[\text{success} \mid w] &\approx (1-2^{-w})^{m_1+1}(1-2^{-d})^{m_1}\\ &\approx (1-2^{-w})^{m_1}(1-2^{-d})^{m_1}.\end{aligned}$$

Next, we argue that the dependence on d can be neglected. Indeed, the constraints in Eq. (10) enforce $d \geq w$, so that the success probability differs significantly from 1 if and only if w is small. Now recall from Eq. (9) that, conditioned on w, the value of d is sampled from a binomial distribution with expectation value

$$\mathbb{E}[d \mid w] \approx \frac{3}{4}\frac{m_1-g}{2} = \frac{3}{8}w + \frac{3}{8}(g+m-n)$$

and then postselected to satisfy $n-g \geq d \geq w$.

For $(n,m,g) = (300,360,4)$ and $w \leq 20$, the expectation value $\mathbb{E}[d \mid w] \approx 0.375w + 20$ is sufficiently far away from the boundaries imposed by the constraint that the effects of the postselection may be neglected. Then, in this parameter regime, we expect $d \simeq w + 20$, so that $2^{-d} \ll 2^{-w}$.

Hence

$$\begin{aligned}\mathrm{Prob}[\text{success} \mid w] &\approx (1-2^{-w})^{m_1} = (1-2^{-w})^{w+g+m-n}\\ &\approx (1-2^{-w})^{g+m-n}.\end{aligned}$$

The final expression is a sigmoid function that reaches the value 1/2 at

$$w_{1/2} = -\log_2\left(1-2^{1/(g+m-n)}\right) \approx 6.3$$

and we therefore predict that the probability of success of the Radical Attack transitions from 0 to 1 around a value of w between 6 and 7.

B. Numerical experiments for the stabilizer construction

We have sampled $\gtrapprox$ 100 000 instances of $\mathbf{H}$ for $m = 360$, $n = 300$, and $g = 4$ (for the computer code and the raw data, see Ref. [28]). Only in 154 cases did the Radical Attack fail to uncover the secret.

Given the number of approximations made, the theoretical analysis turns out to give a surprisingly accurate quantitative account of the behavior of the attack. This is visualized in Fig. 1. In particular, the transition from expected failure to expected success around $w_{1/2} \simeq 6.3$ can

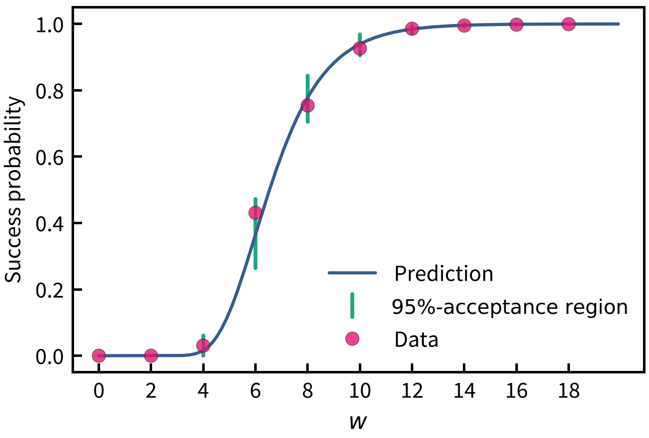


FIG. 1. The probability of success of the Radical Attack given the "excess width" w of the matrix $(\mathbf{B} \mid \mathbf{C})$. The solid sigmoidal curve is the simplified theoretical estimate $\mathrm{Prob}[\text{success} \mid w] \approx (1-2^{-w})^{g+m-n}$. The red dots represent empirical success probabilities for all values of w for which failures have been observed during 100 000 numerical runs. Each vertical bar is the acceptance region of a test for compatibility with the theory prediction at significance level $\alpha = 5\%$. The plot is truncated at $w = 18$, as this is the largest value for which RADICALATTACK() has failed at least once to recover the correct secret in the experiment. The mean value of w is about 32.3 and more than 96% of all instances were associated with a value of w exceeding 18. The simplified theoretical analysis reproduces the behavior of the algorithm in a quantitatively correct way, including predicting the transition from likely failure to likely success at around $w \approx 6$.

be clearly seen in the data. Consistent with the theory, no failures were observed for instances with $w > 18$.

1. Uncertainty quantification for the numerical experiments

Because many of the predicted probabilities are close to 0 or 1, finding a suitable method of uncertainty quantification is not completely trivial.

Commonly, when empirical findings in the sciences are compared to theoretical predictions, one computes a confidence interval with coverage probability $(1-\alpha)$ for the estimated quantity and checks whether the theory prediction lies within that interval. Operationally, this furnishes a statistical hypothesis test for the compatibility between data and theory at significance level α (i.e., the probability that this method will reject data that is in fact compatible is at most α). However, among the set of all hypothesis tests at a given significance level, some are more powerful than others, in the sense that they reject more data sets. The common method just sketched turns out to be of particularly low power in our setting.

Indeed, consider the extreme case in which the hypothesis is $X \sim \mathrm{Binom}(N, p=0)$. Then, a single instance of a nonzero outcome $X_i = 1$ is enough to refute the hypothesis at any significance level. In other words, the acceptance region for the empirical probability $\hat{p} = |\{i \mid X_i = 1\}|/N$ is just $\{0\}$. On the other hand, the statistical *rule of three* states that if no successes have been observed in N attempts, a 95%-confidence region needs to have size about $3/N$.

Happily, for testing compatibility with the predicted parameter of a binomial distribution, there is a *uniformly most powerful unbiased (UMPU)* test [29, Chapter 6.2]. The vertical bars in Fig. 1 represent the resulting acceptance region. We reiterate that this test is much more stringent than the more common approach based on confidence intervals would be.

C. The challenge data set

In light of the very high success rate observed on randomly drawn examples, it came as a surprise to us that an initial version of our attack *failed* for the challenge data set that the authors provided in Ref. [25]. Fortunately, Bremner, Cheng, and Ji were kind enough to publish the full version-control history of their code [25]. The challenge was added with `commit d485f9`. Later, `commit 930fc0` introduced a bug fix in the row-redundancy routine. Under the earlier version, there was a high probability of the $\mathrm{range}(\mathbf{B} \mid \mathbf{C}) = \mathbb{F}_2^{m_2}$ condition failing. In this case, elements of $\ker \mathbf{G}$ would not necessarily correspond to elements of the radical. However, in the challenge data set, the doubly even part of $\mathbf{H}(\ker \mathbf{G})$ *is* contained in the radical. A minimalist fix—removing all singly even columns from the generator matrix for $\mathbf{H}(\ker \mathbf{G})$—suffices to recover the hidden parameters:

$$g = 4, \quad d = 35, \quad m_1 = 96,$$

and the secret

$$\mathbf{s} = \texttt{cyCxfXKxLxXu3YWND2fSzf} \\ \texttt{+YKtZJFLWY1J0l2rBao0A5zVWRSKA=},$$

given here as a base64-encoded binary number. The string has since been kindly confirmed by Bremner, Cheng, and Ji as being equal to the original secret.

D. Application to the Extended Quadratic Residue Code construction

The Radical Attack performs even better against the QRC construction with parameters

$$q \in \{103, 127, 151, 167, 223\}, \quad r = \frac{q+1}{2}, \\ m = 2q, \quad n = r + q$$

recommended by Bremner, Cheng, and Ji as most resilient against the Kahanamoku-Meyer approach (see Sec. III C). In 20 000 runs, we have found not a single instance in which the Radical Attack fails for these parameter choices (for the code and raw data, see Ref. [28]).

The Extended QRC construction does not fix n to $r+q$ but, rather, allows for all values between r and $q+r$. This raises the possibility that there is a parameter regime in which both the Linearity Attack and the Radical Attack fail. We explore this possibility in Fig. 2. We find that the Radical Attack succeeds with high probability for $n \gtrsim q + 13$. As discussed after Corollary 1, this matches the regime in which we expect

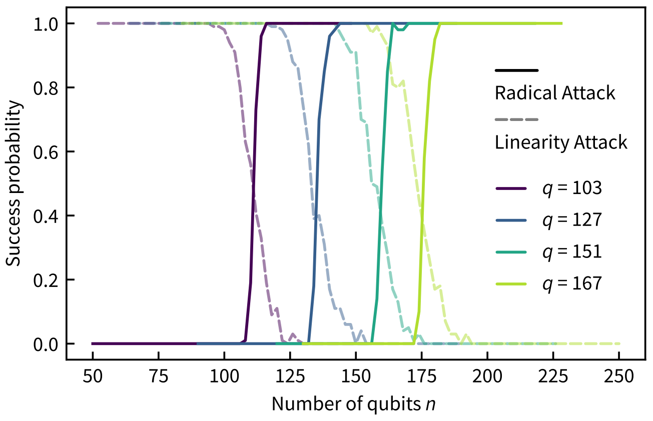


FIG. 2. The performance of the Radical Attack and the Linearity Attack [23] on the updated QRC construction of Bremner, Cheng, and Ji [24], using 100 random instances per point. The Linearity-Attack data are from Ref. [24, Fig. 3b]. The two approaches are seen to complement each other almost perfectly.

(1) the rank of the added row redundancy to saturate such that the condition $\mathrm{range}(\mathbf{B}|\mathbf{C}) = \mathbb{F}_2^{m_2}$ of Corollary 1 is satisfied and
(2) the parameter $w = n - q - 1$ to exceed $\log_2 m_1 = \log_2 q$, which for the choices of $q = 103, 127, 151, 167$ is given by $\log_2 q \approx 7$

Let us also note that the QRC code is guaranteed to have the $\mathbf{1}_{[m_1]} = \mathbf{1}_{[q]}$ vector as a code word and, hence, it is guaranteed to have full support on the obfuscated coordinates.

Comparing this performance with the Linearity Attack, we find only a very slim region around $n \approx q + 13$ in which there exist instances that cannot be solved by either attack with near certainty. This motivates the exploration of further cryptanalytic approaches in Sec. V, where we will indeed present two algorithms—the *Lazy Linearity Attack* and the *Double Meyer*, both building on the approach of Kahanamoku-Meyer [23]—that will eliminate the remaining gap.

V. FURTHER ATTACKS

The IQP Stabilizer Scheme features a large number of degrees of freedom that may allow an algorithm designer to evade any given exploit. The purpose of this section is to exhibit a variety of further approaches that might aid the cryptanalysis of obfuscated IQP circuits. Because the goal is to give an attacker a wide set of tools that may be adapted to any particular future construction, we focus on breadth and, as compared to Sec. IV, put less emphasis on rigorous arguments.

A. The Lazy Linearity Attack

We begin by slightly extending the Linearity Attack to what we call the *Lazy Linearity Attack*, summarized in Algorithm 2. In addition to the IQP tableau $\mathbf{H}$, this routine requires additional input parameters that we call the *ambition A*, the *endurance E*, and the *significance threshold* g_{th}.

1. Analysis

We start by briefly recapitulating why the Linearity Attack of Kahanamoku-Meyer [23] is effective. Essentially, it is based on the following property of the kernel of the Gram matrix $\mathbf{G}_\mathbf{d}$ for vectors $\mathbf{d} \in \mathbb{F}_2^n$.

Lemma 3. For $\mathbf{d} \in \mathbb{F}_2^n$, let $\mathbf{G}_\mathbf{d} := \mathbf{H}_\mathbf{d}^T \mathbf{H}_\mathbf{d}$. The following implication is true:

$$\mathbf{H}_\mathbf{s}\mathbf{d} \in \mathrm{rad}\,\mathcal{C}_\mathbf{s} \quad \Rightarrow \quad \mathbf{s} \in \ker(\mathbf{G}_\mathbf{d}). \tag{15}$$

ALGORITHM 2. Lazy Linearity Attack.

```
function LazyLinearityAttack($\mathbf{H}, g_{\mathrm{th}}, A, E$)
    while $\epsilon < E$ do
        Draw a uniformly random $\mathbf{d} \leftarrow \mathbb{F}_2^n$
        $\mathbf{G}_\mathbf{d} \leftarrow \mathbf{H}_\mathbf{d}^T \mathbf{H}_\mathbf{d}$
        if $\dim \ker \mathbf{G}_\mathbf{d} < A$ then
            for $\mathbf{x} \in \ker \mathbf{G}_\mathbf{d}$ do
                if $\mathrm{rad}\,\mathrm{range}(\mathbf{H}_\mathbf{x}) \neq \{0\}$ and doubly-even and $\mathrm{rank}(\mathbf{H}_\mathbf{x}^T \mathbf{H}_\mathbf{x}) \leq g_{\mathrm{th}}$ then
                    return $\mathbf{x}$ and exit.
                end if
            end for
        end if
        $\epsilon \leftarrow \epsilon + 1$
    end while
    return "fail"
end function
```

Proof. By definition, it holds that $\mathbf{H_s d} \in \operatorname{rad} \mathcal{C}_\mathbf{s}$ if and only if

$$\mathbf{d}^T \mathbf{H}_\mathbf{s}^T \mathbf{H}_\mathbf{s} \mathbf{e} = \mathbf{d}^T \mathbf{H}^T \mathbf{H}_\mathbf{s} \mathbf{e} = 0 \quad \forall\, \mathbf{e} \in \mathbb{F}_2^n. \tag{16}$$

Because $\mathbf{s} \mapsto \mathbf{H_s}$ is linear, the above means that every element $\mathbf{Hd} \in \operatorname{rad} \mathcal{C}_\mathbf{s}$ of the radical gives rise to a set of linear equations (one for each $\mathbf{e} \in \mathbb{F}_2^n$) for the secret $\mathbf{s}$. These equations can be compactly written as

$$\mathbf{d}^T \mathbf{H}^T \mathbf{H}_\mathbf{s} = 0 \quad \Leftrightarrow \quad \mathbf{H}^T \mathbf{H}_\mathbf{d} \mathbf{s} = 0 \quad \Leftrightarrow \quad \mathbf{H}_\mathbf{d}^T \mathbf{H}_\mathbf{d} \mathbf{s} = 0. \tag{17}$$

■

In the Linearity Attack, the strategy is now to pick $\mathbf{d}$ at random. Then, with probability

$$\frac{|\operatorname{rad} \mathcal{C}_\mathbf{s}|}{|\mathcal{C}_\mathbf{s}|} = \frac{2^{\dim \operatorname{rad} \mathcal{C}_\mathbf{s}}}{2^{\dim \mathcal{C}_\mathbf{s}}} = 2^{-g}, \tag{18}$$

$\mathbf{d}$ lies in the radical of $\mathcal{C}_\mathbf{s}$ and we get a constraint on $\mathbf{s}$. If the kernel of $\mathbf{G_d}$ is typically small, one can iterate through all candidates for $\mathcal{C}_\mathbf{s}$ and check the properties of the true $\mathcal{C}_\mathbf{s}$, namely, that $\operatorname{rad} \mathcal{C}_\mathbf{s}$ is doubly even and that its is given by g.

Since the rows of $\mathbf{H}$ are essentially random, we expect that $\mathbf{H_d}$ has around $m/2$ rows that are linearly independent. In the original scheme of Shepherd and Bremner [19], $n \simeq m/2$, and we thus expect $\dim \ker(\mathbf{G_d}) \in O(1)$. More precisely, Bremner, Cheng, and Ji show that indeed $\mathbb{E}_\mathbf{d}[\dim \ker \mathbf{G_d}] \geq n - m/2$. Thus, the running time of the Linearity Attack scales exponentially with $n - m/2$. In the new challenge of Bremner, Cheng, and Ji, $n = 300$ and $m = 360$, so that $n - m/2 = 120$, meaning that $\ker \mathbf{G_d}$ is so large that this simple approach is no longer feasible.

In fact, for $n - m/2 > 0$, the kernel of $\mathbf{H_d}$ (which is contained in $\ker \mathbf{G_d}$) will already be nontrivial. But the relevant part of the kernel of $\mathbf{G_d}$ in which the secret is hiding is independent of n so long as $g + d < n$ and only requires that $\mathbf{d}$ has zero entries in the obfuscated coordinates of $\mathbf{F}$. Thus, we expect that $\dim \ker \mathbf{G_d}$ is roughly independent of the event $\mathbf{s} \in \ker \mathbf{G_d}$. We can thus allow ourselves to ignore large kernels in the search for $\mathbf{s}$, if we are less ambitious about exploring those very large kernels, boosting the success probability.

More precisely, we expect that $\langle \mathbf{H}_i, \mathbf{d} \rangle = 1$ with probability $1/2$. Thus, the number of rows $m_\mathbf{d}$ of $\mathbf{H_d}$ will follow a binomial distribution

$$m_\mathbf{d} \sim \operatorname{Bin}\left(N \approx m, p = 0.5\right), \tag{19}$$

with mean $m/2$ and standard deviation $\sqrt{m}/2$. Since most rows of $\mathbf{H}$ are linearly independent and we expect the values of $\langle \mathbf{H}_i, \mathbf{d} \rangle$ to be only weakly correlated, we thus expect the dimension of the kernel to be given by $\dim(\ker \mathbf{H_d}) \approx n - m_\mathbf{d}$, which is roughly Gaussian around $n - m/2$ with standard deviation $\sqrt{m}/2$ (for a more precise statement, see Ref. [26, Theorem 3.2.2]).

For the Lazy Linearity Attack, the relevant parameter determining the success of the attack is the probability of observing a small kernel in the tail of the distribution over kernels of $\mathbf{H_d}$, induced by the random choice of $\mathbf{d}$. As discussed above, this probability decreases exponentially with $n - m/2 = g + w + i$, where we have defined the *imbalance*

$$i := \frac{m_2 - m_1}{2}. \tag{20}$$

Let the cumulative distribution function of the Gaussian distribution with mean μ and standard deviation σ be given by $C_{\mu,\sigma} : \mathbb{R} \to [0, 1]$. Then, the expected endurance required for the Lazy Linearity Attack with ambition A to succeed is given by

$$E \sim \frac{2^g}{C_{n-m/2,\sqrt{m}/2}(A)}. \tag{21}$$

In numerical experiments, we find our predictions to be accurate up to a constant offset in the predicted mean of $\dim(\ker \mathbf{H})$ [see Fig. 3(a)]. In particular, the dimension of $\ker \mathbf{G} / \ker \mathbf{H}$ is independent of n, which is indeed evidence that there is no correlation between the size of the kernel of $\mathbf{G}$ and whether or not it contains a secret.

2. Application to the Extended QRC construction

Applying the Lazy Linearity Attack to the Extended QRC construction, we find that it succeeds with a near-unit success rate until $n - m/2 \sim 10$, with parameters $A = 8$ and $E = 1000$. Combined with the Radical Attack, we are now able to retrieve the secrets for all proposed parameters of the Extended QRC construction (see Fig. 4). For the Extended QRC construction, $m_1 = m_2$, which means that $i = 0$ and hence the success probability of the Lazy Linearity Attack decreases exactly with $w + g$ and the tunable parameters of the method.

B. The Double Meyer Attack

In the previous section, we have discussed how to exploit statistical fluctuations to avoid having to search through large kernels. But as $m - n/2$ increases, this strategy will eventually fail to be effective. We thus present another ansatz, the *Double Meyer Attack*, stated as Algorithm 3. It reduces the size of the kernel, essentially by running several Linearity Attacks at once.

1. Analysis

At a high level, the Double Meyer examines the elements of the intersection of the kernels of several $\mathbf{G}_{\mathbf{d}^i}$,

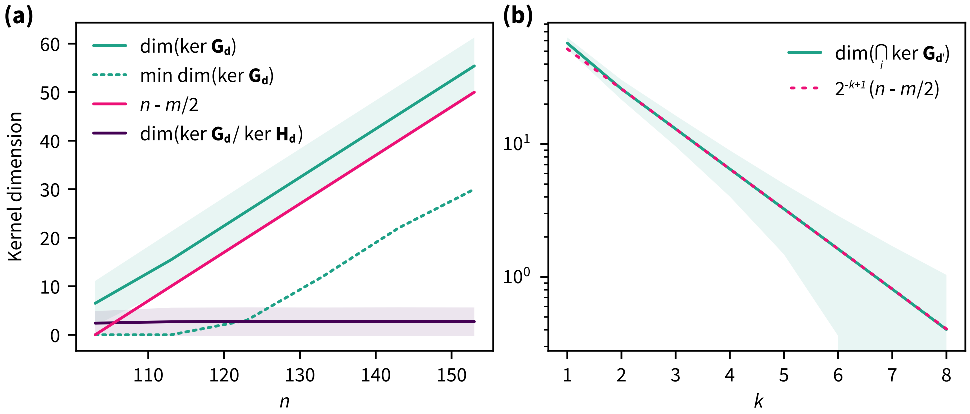


FIG. 3. (a) The dimension of $\ker(\mathbf{G_d})$ (green) for the QRC construction with $q = 103$ and $m = 2q$ for 100 random instances and 1000 random choices of $\mathbf{d}$ per point. The shaded areas represent one standard deviation. This is compared to the simplified theoretical prediction $n - m/2$ (pink). The dotted line designates the minimum observed value of $\dim(\ker \mathbf{G_d})$. The original Linear Attack runs in time roughly exponential in the green curve, whereas the "lazy" approach reduces this to about the exponential of the dotted one. Finally, the violet line depicts $\ker(\mathbf{G_d})/\ker(\mathbf{H_d})$. The fact that it does not depend on n is compatible with the expectation that the probability of finding the secret $\mathbf{s}$ in the kernel of any given Gram matrix $\mathbf{G_d}$ is roughly independent of the size of the kernel. (b) The dimension of $\bigcap_{i\in[k]} \ker(\mathbf{G_{d^i}})$ (green) for the QRC construction with $q = 103, n = q + r, r = (q + 1)/2, m = 2q$. We have used 100 random instances and 1000 random choices of $\mathbf{d}^1, \ldots, \mathbf{d}^k$ per point. The simple theoretical prediction of $2^{-k+1}(n - m/2)$ (dotted pink) is seen to be in good agreement with the numerical experiments.

where the $\mathbf{d}^i$ are uniformly distributed random vectors [30]. Since the events $\mathbf{H_s d}^i \in \operatorname{rad} \mathcal{C}_\mathbf{s}$ are independent for different i's, we have that

$$\Pr\left[\mathbf{s} \in \bigcap_{i\in[k]} \ker \mathbf{G_{d^i}}\right] = (2^{-g})^k = 2^{-gk}. \qquad (22)$$

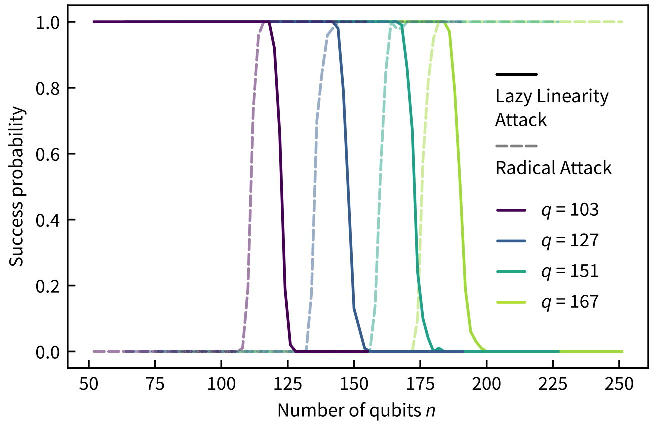


FIG. 4. The performance of the Lazy Linearity Attack (solid lines) with ambition $A = 8$, endurance $E = 1000$, and significance threshold $g_{\text{th}} = 1$, and the Radical Attack (dashed lines) for values of $q = 103$, 127, 151, and 167, with 100 instances per point. Compared to Fig. 2, we find that the "lazy" approach has extended the range of n for which the Linear Attack recovers the secret with near certainty. The shift is sufficient that the two algorithms now cover the entire parameter range.

At the same time, we expect $\dim(\bigcap_{i\in[k]} \ker \mathbf{G_{d^i}}) \approx 2^{-k+1}(m - n/2)$. To see this, observe that the kernel of $\mathbf{G_{d^i}}$ decomposes as $\ker \mathbf{G_{d^i}} = \ker \mathbf{H_{d^i}} + \operatorname{rad} \operatorname{range}(\mathbf{H_{d^i}})$. For random vectors $\mathbf{d}^i$, we then expect $\operatorname{rad} \operatorname{range}(\mathbf{H_{d^i}})$ to be independent identically distributed random subspaces, while the kernels of $\mathbf{H_{d^i}}$ are correlated, since every pair of $\ker \mathbf{H_{d^1}}, \ker \mathbf{H_{d^2}}$ shares half the rows. Thus, the intersection of $\operatorname{rad} \operatorname{range} \mathbf{H_{d^i}}$ decays exponentially with k, while the intersection $\bigcap_{i\in[k]} \ker \mathbf{H_{d^i}}$ decreases much faster, as $n - (1 - 2^{-k})m$. Altogether, the intersection decomposes into sums of intersections (for $k = 2$)

$$\bigcap_{i=1}^{2} \ker \mathbf{G_{d^i}} = (\ker \mathbf{H_{d^1}} + \operatorname{rad} \operatorname{range}(\mathbf{H_{d^1}})) \cap (\ker \mathbf{H_{d^2}} + \operatorname{rad} \operatorname{range}(\mathbf{H_{d^2}})), \qquad (23)$$

where the exponential decay as 2^{-k} stemming from the last term dominates the scaling.

Choosing $k \sim \log n$ is sufficient to reduce the kernel dimension to $O(1)$. Moreover, g needs to be of order $O(\log n)$ in order to maintain a sample-efficient verification test for the challenge. Thus, $2^{gk} \in 2^{O(\log^2(n))}$, i.e., the Double Meyer Attack is expected to run in at most quasipolynomial time, for any choice of n, m. However, even moderate values of g will make this approach infeasible in practice. For the challenge data set, we expect that good parameter choices are $k = 6, A = 3$, and $E \gtrsim 2^{gk} = 2^{24} \approx 10^{7.22}$, ALGORITHM 3. Double Meyer Attack.

```
function DoubleMeyer($\mathbf{H}, k, g_{\text{th}}, A, E$)
    while $\epsilon < E$ do
        for $i \in [k]$ do
            Draw a uniformly random $\mathbf{d}^i \leftarrow \mathbb{F}_2^n$
            $\mathbf{G}_{\mathbf{d}^i} \leftarrow \mathbf{H}_{\mathbf{d}^i}^T \mathbf{H}_{\mathbf{d}^i}$
            $\mathbf{G} \leftarrow (\mathbf{G}^T | \mathbf{G}_{\mathbf{d}^i}^T)^T$
        end for
        if $\dim \ker \mathbf{G} < A$ then
            for $\mathbf{x} \in \ker \mathbf{G}_{\mathbf{d}}$ do
                if $\operatorname{rad} \operatorname{range}(\mathbf{H}_{\mathbf{x}}) \neq \{0\}$ and doubly-even and $\operatorname{rank}(\mathbf{H}_{\mathbf{x}}^T \mathbf{H}_{\mathbf{x}}) \leq g_{\text{th}}$ then
                    return $\mathbf{x}$ and exit.
                end if
            end for
        end if
        $\epsilon \leftarrow \epsilon + 1$
    end while
    return “fail”
end function
```

though we have not spent enough computational resources to have recovered a secret in this regime.

We observe that slack between the threshold rank g_{th} and the true g often leads to a misidentified secret. This is explained by vectors $\mathbf{v} \in \bigcap_{i \in [k]} \ker \mathbf{H}_{\mathbf{d}^i}$, the image of which under $\mathbf{H}$ has low Hamming weight $|\mathbf{H}\mathbf{v}| \leq g_{\text{th}} - g$. This corresponds to rows of $\mathbf{H}$ that can be mapped to $s = g_{\text{th}} - g$ unit vectors $\mathbf{e}^1, \dots, \mathbf{e}^s$ that are linearly independent of the first m_1 rows of $\mathbf{H}$. These vectors may be absorbed into $\mathbf{F}$, adding s nontrivial columns to it, while adding a zero row to $\mathbf{D}$, which keeps its range doubly even. The alternative secrets $\mathbf{v}$ found in this way are also observable in the sense that they satisfy $\Pr_{\mathbf{x} \leftarrow D_{\mathbf{H}}}[\langle \mathbf{x}, \mathbf{v} \rangle = 0] = (2^{-(g+s)/2} + 1)/2$. In order to find the “true” secret, one should therefore run the attack for increasing values of g_{th} and halt as soon as a valid secret is found.

The low-Hamming-weight vectors identified above inform the final ansatz of this paper, *Hamming’s Razor*, presented in Sec. V C.

2. Application to the Extended QRC construction

We find that the Double Meyer Attack recovers the secrets of the Extended QRC construction with near certainty in all parameter regimes proposed by Bremner, Cheng, and Ji. For the QRC construction, $g = 1$ and hence the running time of the Double Meyer Attack is just given by 2^k. Choosing $k = 6$, $g_{\text{th}} = 1$, $E = 8$, and $A = 1000$ is sufficient to recover the secret in all of 100 random instances of the Extended QRC construction for all values of $n \in [r, q + r]$ and $q \in \{103, 127, 151, 167\}$.

3. Further improvements

As stated, the Double Meyer draws vectors $\mathbf{d}^i$ uniformly at random, in the hope that $\mathbf{H}_{\mathbf{s}}\mathbf{d}^i \in \operatorname{rad} \mathcal{C}_{\mathbf{s}}$. But there might be more efficient ways of obtaining vectors $\mathbf{d}^i$ satisfying this condition. For instance, under the assumptions of Corollary 1 (1), *any* element $\mathbf{d}$ of $\mathbf{H}(\ker(\mathbf{H}^T\mathbf{H}))$ has this property. Adding such vectors to the collection of instances of $\mathbf{d}^i$ therefore provides additional constraints for $\mathbf{s}$ at essentially no computational cost.

In particular, this modification of the Double Meyer would break a variant of the construction that Bremner *et al.* [31] have proposed as an initial reaction to the preprint of this paper with parameters `-AB-type zero -concat_D -concat_C1` as implemented in `commit 7d3bd3` of their GitHub repository [32].

C. Hamming's Razor

In this section, we describe a method that allows one to "shave off" certain rows and columns from $\mathbf{H}$ without affecting the code space $\mathcal{C}_\mathbf{s}$. Such redundancies can be identified given a vector $\mathbf{d} \in \mathbb{F}_2^n$ such that $\mathbf{Hd}$ has low Hamming weight. The method comes in two varieties: the simpler *Singleton Razor*, discussed first, and the more general *Hamming's Razor* proper.

1. The Singleton Razor

Let us agree to call $i \in [m]$ a *singleton for* $\mathbf{H}$ if there is a solution to $\mathbf{Hd} = \mathbf{e}^i$.

We discuss the idea based on the unobfuscated picture of Eq. (3). Assume that $\operatorname{supp}\operatorname{range}\mathbf{D} = [m_1]$. Then, there is no singleton among the first m_1 coordinates, because $\mathbf{e}^i$ is not orthogonal to range $\mathbf{D}$, while all vectors in the range of $(\mathbf{F} \mid \mathbf{D})$ are. Thus, knowing a singleton i, one may trim away the ith row of $\mathbf{H}$ without affecting the code space. Alternatively, one can perform a coordinate change on $\mathbb{F}_2^n$ that maps the corresponding preimage $\mathbf{d}$ to $\mathbf{e}^n$ and then drop the nth column of $\mathbf{H}$. In fact, both operations may be combined, without changing $\mathcal{C}_\mathbf{s}$.

The generator matrix $\mathbf{H}$ of the challenge data set affords 69 singletons. All singletons do indeed belong to redundant rows [28]. The second part of Lemma 4 below suggests an explanation for this surprisingly high number.

2. Hamming's Razor

We now generalize the singleton idea to higher Hamming weights. The starting point is the observation that range $\mathbf{C}$ can be expected to contain vectors of much lower Hamming weight than range $(\mathbf{F} \mid \mathbf{D})$. This will lead to a computationally efficient means for separating redundant from nonredundant rows.

The following lemma collects two technical preparations.

Lemma 4. Let $\mathbf{M}$ be an $m \times (m-h)$ binary matrix chosen uniformly at random. The probability that the minimal Hamming weight of any nonzero vector in the range of $\mathbf{M}$ is smaller than k is exponentially small in $k_1 - k$, where

$$k_1 = h\frac{\ln 2}{\ln m + 2} \approx \frac{h}{\log_2 m}.$$

More precisely and more strongly, the probability is no larger than $e^{-\lambda(k_\infty - k)}$, for

$$k_\infty = \lim_{i\to\infty} k_i, \quad k_i = k_1 + \frac{(k_{i-1}-1)\ln k_{i-1}}{\ln m + 2},$$
$$\lambda = \frac{1}{k_\infty} + \ln\frac{m}{k_\infty} > 0. \tag{24}$$

Conversely, let $\mathbf{M}$ be any binary matrix the range of which has co-dimension h. For $S \subset [m]$, let $V_S \subset \mathbb{F}_2^m$ be the subspace of vectors supported on S. Then, range M has a nontrivial intersection with V_S if $|S| > h$.

Proof. Let $\mathbf{v}$ be a random vector distributed uniformly in $\mathbb{F}_2^m$. Then, as long as $k \le m/2$,

$$\Pr[|\mathbf{v}| \le k] = \sum_{k' \le k}\binom{m}{k'}2^{-m} \le k\binom{m}{k}2^{-m} \quad \Rightarrow$$
$$\Pr\left[\min_{0\neq\mathbf{v}\in\operatorname{range}\mathbf{M}} |\mathbf{v}| \le k\right] \le k\binom{m}{k}2^{-h}.$$

Using the standard estimate $\ln\binom{m}{k} \le k(\ln m + 1) - k\ln k$, the logarithm of the bound is

$$l(k) := k(\ln m + 1) - (k-1)\ln k - h\ln 2.$$

The function $l(k)$ is concave, negative at $k = 0$, positive at $k = m/2$, and thus has a zero in the interval $[0, m/2]$. What is more, $l(k) = 0$ if and only if

$$k = h\frac{\ln 2}{\ln m + 1} + \frac{(k-1)\ln k}{\ln m + 1}.$$

Because the right-hand side is monotonous in k, the recursive formula in Eq. (24) defines an increasing sequence k_i of lower bounds to the first zero. As a nondecreasing sequence on a bounded set, the limit point k_∞ is well defined. Due to concavity, $l(k)$ is upper bounded by its first-order Taylor approximations. The claim then follows by expanding around $k = k_\infty$.

The converse statement is a consequence of the standard estimate

$$\dim(\operatorname{range}\mathbf{M} \cap V_S) \ge \dim\operatorname{range}\mathbf{M} + \dim V_S - m.$$

■

Choose a set $S \subset [m]$ and let $\mathbf{H}_{\setminus S}$ be the matrix obtained by deleting all rows of $\mathbf{H}$ the index of which appears in S. Let $S_1 = S \cap [m_1]$ be the intersection of S with the "secret rows" and let $S_2 = S \cap [m_1 + 1, m_2]$ be the intersection with the redundant rows. Then, $\mathbf{d} \in \ker\mathbf{H}_{\setminus S}$ if and only if

$$\operatorname{supp}(\mathbf{H}_\mathbf{s}\mathbf{d}) \subset S_1 \quad \text{and} \quad \operatorname{supp}(\mathbf{R}_\mathbf{s}\mathbf{d}) \subset S_2.$$

Now model $(\mathbf{F} \mid \mathbf{D})$ as a uniformly random matrix. Lemma 4 applies with $m = m_1$ and $h = m_1 - g - d$, giving rise to

ALGORITHM 4. Hamming's Razor.

```
function HammingRazor($\mathbf{H}, p, E$)
    $S \leftarrow \emptyset$.
    for $\_ \in [E]$ do
        Draw a random vector $\mathbf{d} \in \mathbb{F}_2^m$ with entries $\mathbf{d}_i \leftarrow \mathrm{Bin}(\mathbb{F}_2, p)$.
        $\mathbf{H}[\mathbf{d}] \leftarrow \mathrm{diag}(\mathbf{d})\mathbf{H}$.
        $\mathbf{K} \leftarrow$ a column-generating matrix for $\ker \mathbf{H}[\mathbf{d}]$.
        Append the support of the columns of $\mathbf{HK}$ to $S$.
    end for
    Solve the $\mathbb{F}_2$-linear system $\mathbf{Hs} = \mathbf{1}_{S^c}$
    return $\mathbf{s}$
end function
```

an associated value of k_∞. If $|S_1| < k_\infty$, then, up to an exponentially small probability of failure, the first condition can be satisfied only if $\mathbf{H_s d} = 0$. Therefore, each nonzero element $\mathbf{d} \in \ker \mathbf{H}_{\setminus S}$ identifies $\mathrm{supp}(\mathbf{Hd})$ as a set of redundant rows, which can be eliminated as argued in the context of the Singleton Attack.

This observation is useful only if it is easily possible to identify suitable sets S and nonzero vectors $\mathbf{d}$ in the associated kernel. Here, the second part of Lemma 4 comes into play. If $|S_2| > m_2 - (n - g - d)$, then by the lemma and Eq. (8), there exists a nonzero $\mathbf{d}$ such that $\mathrm{supp}(\mathbf{Cd}) \subset S_2$.

This suggests that we should construct S by including each coordinate $i \in [m]$ with probability p, chosen such that $pm_2 > m_2 - (n - g - d)$ and $pm_1 < k_\infty$. For the challenge parameters, these requirements are compatible with the range $[0.01, 0.13]$ for p.

In fact, one can base a full secret-extraction method on this idea (see Algorithm 4). Repeating the procedure for a few dozen random S turns out to reveal the entire redundant row set, and thus the secret, for the challenge data [28]. The attack may be sped up by realizing that the condition on $|S_1|$ has been chosen conservatively. The first part of Lemma 4 states that the *smallest* Hamming weight that occurs in the range of $\mathbf{H_s}$ is about k_∞. But a *randomly chosen* S_1 of size larger than k_∞ is unlikely to be the support of a vector in range $\mathbf{H_s}$ unless $|S_1|$ gets close to the much larger second bound in the lemma. This optimization, for a heuristically chosen value of $p = 0.25$, is used in the sample implementation provided with this paper [28] and recovers the secret with high probability.

VI. DISCUSSION AND CONCLUSION

In this work, we have exhibited a number of approaches that can be used to recover secrets hidden in obfuscated IQP circuits.

As a reaction to a preprint version of this paper, Bremner, Cheng, and Ji have modified their proposal to evade the attacks described here. A first update (communicated privately) led to our improved Double Meyer, as sketched in Sec. V B 3. As of September 2024, the authors have provided us with a version of their protocol in which we have not found a weakness [33].

It may be instructive to compare the situation to the more mature field of classical cryptography, which benefits from a large public record of cryptographic constructions and their cryptanalysis. New protocols can thus be designed to resist known exploits and be vetted against them. In particular, most cryptographic protocols actually in use are not rigorously proven to be secure. Instead, trust in them is based on a long and public history of constructions, as well as successful and unsuccessful attempts at attacking them. The well-documented story of *differential cryptanalysis* provides an instructive example.

In this light, we consider the high-level contributions of our paper to be the following. (1) It shows potential users of crytpographically backed-up demonstrations of quantum advantage that previous proposals have been broken repeatedly, so that their security should not be taken for granted. (2) The fact that we have not yet been able to identify an efficient attack against the latest version of the protocol should raise one's trust in that version of the proposal, compared to a protocol that has not been the subject of a public security review. (3) Our attacks clarify properties of the IQP-based protocol that make it amenable to classical attacks as well as a collection of cryptanalytic techniques exploiting those properties. These tools can guide and must be taken into account by designers of future constructions.

At the same time, our cryptanalysis still has two important consequences on IQP-based verified quantum advantage using the construction of Bremner *et al.* [24]. First, our *Double Meyer* attack remains valid for all instances and has quasipolynomial running time $2^{-\Omega(\log^2 n)}$, given the verification condition that the signal 2^{-g} remains inverse-polynomially large. This is a significant improvement over the previous state of the art, which was an exponential-time algorithm. Second, in order to circumvent our attacks, Bremner *et al.* [33] have had to significantly increase the problem or “key” size from the original challenge data with parameters $n = 300$, $m = 360$, and $g = 4$ to $n = 700$ qubits, $m = 1200$ Hamiltonian terms, and a verification signal of $g = 10$. It is therefore an interesting open problem to compare the implementation cost of this scheme to other resource-efficient schemes that come with provable security guarantees based on well-studied classical assumptions [5,6].

We emphasize that, in our judgment, the problem of finding ways to efficiently certify the operation of near-term quantum computing devices is an important one and the idea of using obfuscated quantum circuits remains appealing. More generally, the story of IQP-based verified quantum advantage and our contribution to it illustrates that results that exhibit weaknesses in published constructions should not cause the community to turn away but, rather, should serve as sign posts guiding the way to more resilient schemes. We remain curious whether the security of the new construction of Bremner *et al.* [24] holds up to further scrutiny by the community.

ACKNOWLEDGMENTS

First and foremost, we are grateful to Michael Bremner, Bin Cheng, and Zhengfeng Ji for setting the challenge and providing a thorough code base that set a low threshold for its study. We feel it is worth stating that we have tried numerous times to construct alternative hiding schemes ourselves—but found each of our own methods much easier to break than theirs. We also thank Mick, Bin, and Zhengfeng for detailed comments and discussions on earlier versions of this paper. We are grateful for many inspiring discussions on hiding secrets in a circuit description, with Dolev Bluvstein, Alex Grilo, Michael Gullans, Jonas Helsen, Marcin Kalinowski, Misha Lukin, Robbie King, Yi-Kai Liu, Alex Poremba, and Joseph Slote. D.G.’s contribution has been supported by the German Ministry for Education and Research under the QSolid project, and by Germany’s Excellence Strategy—Cluster of Excellence Matter and Light for Quantum Computing (ML4Q) EXC 2004/1 (390534769). D.H. acknowledges funding from the U.S. Department of Defense through a QuICS Hartree fellowship and from the Simons Institute for the Theory of Computing, supported by DOE QSA.

APPENDIX: IMPLEMENTATION DETAILS

While running the software package provided with Ref. [24] tens of thousands of times, we found a number of extremely rare edge cases that were not explicitly handled.

In particular, the `sample_parameters()` and the `sample_D()` functions would very rarely return inconsistent results. Rather straightforward corrections are published in Ref. [28].

Another possible discrepancy between the procedure described in the main text of Ref. [24] and their software implementation concerns the generation of the “redundant rows.” The issue is a little more subtle than the first two, so we briefly comment on it here.

In the paper, the relevant quote is

> “Therefore, up to row permutations, the first $n - r$ rows of $\mathbf{R_s}$ are sampled to be *random independent* rows that are orthogonal to $\mathbf{s}$ and lie outside the row space of $\mathbf{H_s}$”

(emphasis ours). The computer implementation is given by the `add_row_redundancy()` function in `lib/construction.py`—in particular, by these lines:

```
s_null_space = s.reshape((1, -1)).
 null_space()

full_basis = row_space_H_s
for p in s_null_space:
 if not check_element(full_basis.T, p):
 full_basis = np.concatenate
    ((full_basis, p.reshape(1, -1)),
    axis=0)

R_s = full_basis[r:]
    # guarantee that rank(H) = n
```

At this point, the first $n - r$ rows of $\mathbf{R_s}$ are not “random independent.” The behavior of this piece of code depends on the detailed implementation of the `null_space()` function, which we do not directly control. While the obfuscation process will later add randomness, we caution that the *same* random invertible matrix $\mathbf{Q}$ acts both on $\mathbf{R_s}$ and on $\mathbf{H_s}$. Any relation between these two blocks that is invariant under right multiplication by an invertible matrix will therefore be preserved. As a mitigation of this possible effect, we suggest adding an explicit randomization step, such as

```
s_null_space=rand_inv_mat
    (s_null_space.shape[0],seed=rng)
    @s_null_space
```

to the routine (though in practice, we did not observe different behavior between these two versions).

The code published in Ref. [28] contains three implementations of the `add_row_redundancy()` function. The version used to create the challenge data (cf. Sec. IV C), the one published with Ref. [24], and finally the one with the explicit extra randomization step added. The numerical results reported in the main text of this paper were generated by the third routine, though we have also include 20 000 runs performed with the second version [28]. The effectiveness of the Radical Attack does not seem to differ appreciatively between these two implementations.

[1] C. Gidney and M. Ekerå, How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits, Quantum **5**, 433 (2021).

[2] D. Litinski, arXiv:2306.08585.

[3] Z. Brakerski, P. Christiano, U. Mahadev, U. Vazirani, and T. Vidick, in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, Paris, France, 2018), pp. 320–331.

[4] Z. Brakerski, V. Koppula, U. Vazirani, and T. Vidick, in *15th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2020)*. Leibniz International Proceedings in Informatics (LIPIcs) (2020), Vol. 158, pp. 8:1–8:14.

[5] G. D. Kahanamoku-Meyer, S. Choi, U. V. Vazirani, and N. Y. Yao, Classically verifiable quantum advantage from a computational Bell test, Nat. Phys. **18**, 918 (2022).

[6] D. Zhu, G. D. Kahanamoku-Meyer, L. Lewis, C. Noel, O. Katz, B. Harraz, Q. Wang, A. Risinger, L. Feng, D. Biswas, L. Egan, A. Gheorghiu, Y. Nam, T. Vidick, U. Vazirani, N. Y. Yao, M. Cetina, and C. Monroe, Interactive cryptographic proofs of quantumness using mid-circuit measurements, Nat. Phys. **19**, 1725 (2023).

[7] M. J. Bremner, R. Jozsa, and D. J. Shepherd, Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy, Proc. R. Soc. A: Math., Phys. Eng. Sci. **467**, 459 (2010).

[8] S. Aaronson and A. Arkhipov, The computational complexity of linear optics, Theor. Comput. **9**, 143 (2013).

[9] M. J. Bremner, A. Montanaro, and D. J. Shepherd, Average-case complexity versus approximate simulation of commuting quantum computations, Phys. Rev. Lett. **117**, 080501 (2016).

[10] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, Characterizing quantum supremacy in near-term devices, Nat. Phys. **14**, 595 (2018).

[11] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, *et al.*, Quantum supremacy using a programmable superconducting processor, Nature **574**, 505 (2019).

[12] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, *et al.*, Quantum computational advantage using photons, Science **370**, 1460 (2020).

[13] Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, *et al.*, Strong quantum computational advantage using a superconducting quantum processor, Phys. Rev. Lett. **127**, 180501 (2021).

[14] F. Pan, K. Chen, and P. Zhang, Solving the sampling problem of the Sycamore quantum circuits, Phys. Rev. Lett. **129**, 090502 (2022).

[15] G. Kalachev, P. Panteleev, P. Zhou, and M.-H. Yung, arXiv:2112.15083.

[16] A. Morvan, B. Villalonga, X. Mi, S. Mandrà, A. Bengtsson, P. V. Klimov, Z. Chen, S. Hong, C. Erickson, I. K. Drozdov, *et al.*, Phase transitions in random circuit sampling, Nature **634**, 328 (2024).

[17] D. Hangleiter and J. Eisert, Computational advantage of quantum random sampling, Rev. Mod. Phys. **95**, 035001 (2023).

[18] T. Yamakawa and M. Zhandry, Verifiable quantum advantage without structure, J. ACM **71**, 1 (2024).

[19] D. Shepherd and M. J. Bremner, Temporally unstructured quantum computation, Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci. **465**, 1413 (2009).

[20] S. Aaronson, Recent progress in quantum advantage, 2022, talk at the Simons Institute; accessed November 27, 2013, https://simons.berkeley.edu/talks/recent-progress-quantum-advantage.

[21] S. Aaronson, Verifiable quantum supremacy: What I hope will be done, 2023, talk at the Simons Institute; accessed November 27, 2013, https://simons.berkeley.edu/talks/scott-aaronson-university-texas-austin-2023-07-10.

[22] See https://quantumchallenges.wordpress.com for the challenge of Shepherd and Bremner [19].

[23] G. D. Kahanamoku-Meyer, Forging quantum data: Classically defeating an IQP-based quantum test, Quantum **7**, 1107 (2023).

[24] M. J. Bremner, B. Cheng, and Z. Ji, IQP sampling and verifiable quantum advantage: Stabilizer scheme and classical security, arXiv:2308.07152v1.

[25] M. J. Bremner, B. Cheng, and Z. Ji, commit `11d4c52`, 2023, available online at https://github.com/AlaricCheng/stabilizer_protocol_sim.

[26] V. F. Kolchin, *Random Graphs*, Encyclopedia of Mathematics and Its Applications (Cambridge University Press, Cambridge, 1998).

[27] One difficulty to overcome is that it is not apparent that sequential sampling algorithms, such as the one implemented in Ref. [24], produce a uniform distribution over all subspaces compatible with the geometric constraints. Usually, such results are proven by invoking a suitable version of Witt's lemma to establish that any partial generator matrix can be extended to a full one in the same number of ways. In geometries that take Hamming weight modulo 4 into account, there may, however, be obstructions against such extensions. A relevant reference is Ref. [34, Sec. 4] (see also Refs. [35,36] for a discussion in the context of stabilizer theory). The theorem in that section states that two isometric subspaces can be mapped onto each other by a

global isometry only if they both contain the all-ones vector or if neither of them does. The resulting complications have prevented us from finding a simple rigorous version of Lemma 2 that applies to random generator matrices of doubly even spaces.

[28] D. Gross and D. Hangleiter, De-obfuscate IQP, 2023, available online at https://github.com/goliath-klein/deobfuscate-iqp and at https://zenodo.org/records/10407186.

[29] J. Shao, *Mathematical Statistics*, 2nd ed., Springer Texts in Statistics (Springer, New York, 2003).

[30] The essential generalization over Kahanamoku-Meyer's Linearity Attack can be gleaned from the simplest special case, $k = 2$, which, of course, explains the name we have adopted for this ansatz.

[31] M. J. Bremner, B. Cheng, and Z. Ji, IQP Stabilizer Scheme and beyond (private communication).

[32] M. J. Bremner, B. Cheng, and Z. Ji, commit `7d3bd3`, 2024, available online at https://github.com/AlaricCheng/stabilizer_protocol_sim.

[33] M. J. Bremner, B. Cheng, and Z. Ji, Instantaneous quantum polynomial-time sampling and verifiable quantum advantage: Stabilizer scheme and classical security, PRX Quantum **6**, 020315 (2025).

[34] J. A. Wood, Witt's extension theorem for mod four valued quadratic forms, Trans. Amer. Math. Soc. **336**, 445 (1993).

[35] D. Gross, S. Nezami, and M. Walter, Schur-Weyl duality for the Clifford group with applications: Property testing, a robust Hudson theorem, and de Finetti representations, Commun. Math. Phys. **385**, 1325 (2021).

[36] F. Montealegre-Mora and D. Gross, Duality theory for Clifford tensor powers, arXiv:2208.01688.