Quantum obfuscation

Gorjan Alagic and Bill Fefferman

September 19, 2015

Abstract

Encryption of data is fundamental to secure communication in the modern world. Beyond encryption of data lies *obfuscation*, i.e., encryption of functionality. It has been known since 2001 that the most powerful means of obfuscating classical programs, so-called "black-box obfuscation," is provably impossible. For years since, obfuscation was believed to always be either impossible or useless, depending on the particulars of its formal definition. However, several recent results have yielded candidate schemes that satisfy a definition weaker than black-box, and yet still have numerous applications.

In this work, we initialize the rigorous study of obfuscating programs *via quantum-mechanical means.* We define notions of quantum obfuscation which encompass several natural variants. For instance, the input can describe classical or quantum functionality, and the output can be either a classical description or a quantum state. The obfuscator can also satisfy one of a number of obfuscation conditions: black-box, information-theoretic black-box, indistinguishability, and best-possible. We discuss a number of applications, including CPA-secure quantum encryption, quantum fully-homomorphic encryption, and quantum money. We then prove several impossibility results, extending a number of foundational papers on classical obfuscation to the quantum setting. In particular, we prove that black-box obfuscation which outputs quantum circuits is impossible, and that statistical indistinguishability obfuscation which outputs circuits or states is impossible (up to an unlikely complexity-theoretic collapse.) Our proofs involve a new tool: chosen-ciphertext-secure encryption of quantum data, which we show is possible provided that quantum-secure one-way functions exist.

We emphasize that our results leave open one intriguing possibility: obfuscating a classical or quantum circuit into a single, uncloneable quantum state. This indicates that, in spite of our results, quantum obfuscation may be significantly more powerful than its classical counterpart.

Contents

1 Introduction

The ability to encrypt data is central to modern communications. In our daily lives, we frequently make use of a number of powerful tasks in encryption, such as private-key encryption, key exchange, public-key encryption; in the near future, this may even include fully homomorphic encryption. Arguably the most powerful cryptographic ability is *obfuscation*, which enables *encryption of functionality*. Obfuscation implies (with some caveats) the ability to perform almost any other task in encryption; this includes everything in the list above, and much more.

To understand obfuscation, it is useful to think about an obvious application: protecting intellectual property in software. In this setting, a software developer wishes to distribute their software to end users. However, the code undoubtedly contains a number of trade secrets which the developer does not want to become public. In order to accomplish this, the software is first passed through an obfuscator, and then published. The obfuscator must thus be an efficient algorithm that satisfies three core properties:

1. *functional equivalence:* the input/output functionality does not change;
2. *polynomial slowdown:* if the input program is efficient, then the output program is efficient;
3. *obfuscation:* the code of the output program is "hard to understand."

The last condition can be formulated rigorously in a number of ways. One possibility is the so-called "virtual black-box" condition, which says that the obfuscated program is no more useful than an impenetrable box which simply accepts inputs and produces outputs. While this condition appears to be too strong, there are other formulations as well, with varying levels of strength and usefulness.

The above scenario of encrypting classical programs (and classical data) is significantly complicated by the advent of *quantum computation.* One widely-known complication is that certain encryption schemes are no longer secure in the presence of quantum adversaries. The same may hold for obfuscation schemes. On the other hand, quantum mechanics may also enable us to perform cryptographic tasks that are impossible classically. It is thus natural to ask what quantum computation means for obfuscation of programs. In particular, we would like to answer the following questions:

- is it possible to quantumly obfuscate classical programs?
- is it possible to obfuscate quantum programs?
- how should we formulate quantum obfuscation in a rigorous manner?
- which of the classical results about obfuscation carry over to the quantum setting?
- can one use quantum mechanics to obfuscate in ways that are impossible classically?
- are there interesting applications of any of the above?

We remark that, in order to address the above questions, we must also properly address the question of encrypting quantum data—a strictly simpler task than encrypting functionality. While information-theoretic encryption of quantum data has been considered before, in this setting we are interested in encryption of quantum data *with computational assumptions*[1]. This latter subject has not yet received significant attention in literature.

[1]Note that information-theoretic obfuscation is impossible if the adversary can execute the obfuscated program multiple times: a computationally unbounded adversary can then simply evaluate the program on all possible inputs, and use this to learn everything there is to know about the program.

1.1 Background

Things to cover:

- classical work on obfuscation
  - original VBB impossibility result
  - original indistinguishability and best-possible paper
  - recent work on candidate IO schemes, all the nice applications, and a brief summary of latest work
- quantum work on obfuscation
  - Scott's semi-grand challenge question
  - Mosca and Stebila suggestion of obfuscating circuits to get quantum money
  - Scott's claims that were never published
  - our paper on "partial-indistinguishability."

1.2 Summary of results

In this section, we summarize our results. The results are divided by subject, with quantum encryption covered in Section 3, quantum black-box obfuscation in Section 4, and quantum indistinguishability obfuscation in Section 5.

1.2.1 Quantum encryption

For us, *quantum encryption* will mean the encryption of quantum states under computational assumptions. The advantages of this form of encryption over information-theoretic are well-known in the classical setting; for us the crucial advantages will be (i.) reusability of the key, and (ii.) chosen-ciphertext security. The results on quantum encryption which we will present are summarized below, and will be necessary in order to establish some of our results about black-box obfuscation.

1. We define a notion of symmetric-key encryption scheme for quantum states, with reusable keys; these schemes consist of three quantum algorithms (key generation, encryption, and decryption) which satisfy correctness: under a fixed key, encryption followed by decryption must be equivalent to the identity.
2. We define a notion of IND-CCA1 (or *indistinguishability of ciphertexts under non-adaptive chosen ciphertext attacks*) for these schemes; this formalizes the idea of a "lunchtime attack," where an adversary has complete access to all aspects of the encryption except the key itself, and is tasked with decrypting a challenge ciphertext later (presumably after lunch.)
3. We give a construction for an IND-CCA1-secure symmetric-key encryption scheme for quantum states, under the assumption that quantum-secure one-way functions exist. These are deterministic classical functions which are easy to compute, but hard to invert for quantum adversaries.

A number of other contemporaneous works consider quantum encryption with computational assumptions. A complete treatment of the basic notions will appear in [3]. Broadbent and Jeffrey considered IND-CPA-secure public-key and symmetric-key quantum encryption, together with partial homomorphism [7].

1.2.2 Quantum black-box obfuscation

Our main results concern the questions of definitions, applications, and (im)possibility of quantum obfuscation in the virtual black-box setting.

1. define quantum black-box obfuscation
2. show several applications:
   - classical algorithm for it implies quantum-secure OWFs;
   - quantum obfuscator implies IND-CPA SKE for quantum states;
   - obfuscator + qOWFs implies IND-CPA PKE for quantum states;
   - obfuscator (plus what else?) implies QFHE;
   - emphasize that all these encryption schemes also work for *classical data*, but may require quantum ciphertexts (and hence also quantum encryption and decryption algorithms);
   - obfuscator implies quantum money (details?);
3. impossibility results:
   - two-state black-box obfuscation impossible
   - obfuscation with cloneable outputs impossible
4. as part of the impossibility result, we will need the existence of IND-CCA1 secure quantum encryption schemes; so we define those and prove that they exist; this appears to be new as well.

1.2.3 Quantum indistinguishability obfuscation

1. define quantum indistinguishability obfuscation and quantum best-possible obfuscation;
2. three variants: perfect, statistical, computational;
3. applications:
   - witness encryption for QMA
   - classically we also get functional encryption (cite) and many more applications through the very successful "punctured programs" technique (cite); we suspect that these can be adapted to the quantum case, but leave them open for now;
4. proved each indistinguishability variant is equivalent to its corresponding best-possible variant;
5. proved impossibility of perfect and statistical indistinguishability obfuscators.

2 Preliminaries

2.1 Notation etc.

- PT, PPT, QPT, etc. (might be partially covered in the next subsection)
- the notation $x \in_R \{0,1\}^n$.
- $\mathcal{H}_m$ is the space of m-qubit pure states, and $\mathfrak{D}(\mathcal{H}_m)$ is the corresponding space of density operators
- $\mathbb{1}_m$ is the m-qubit identity.

2.2 Probabilistic and quantum algorithms

We briefly review some terminology regarding probabilistic and quantum algorithms. For precise definitions, refer to [**?**]. As is standard, by a probabilistic classical algorithm $\mathcal{A}$ we will mean an infinite family of probabilistic classical circuits, at least one for each possible input size [2]. When the input register is initialized with the string x and the randomness register is initialized with the string r, the output of the relevant circuit will be denoted by $\mathcal{A}(x;r)$. We will simply write $\mathcal{A}(x)$ when the randomness register should be initialized with a uniformly random string.

A quantum algorithm $\mathcal{Q}$ will mean an infinite family of quantum circuits, at least one for each possible input size. For each circuit, the qubits it acts on are divided into an input register and an ancilla register; the former is initialized in some input state σ and the latter is always initialized in the $|0\rangle$ state. All of the qubits are also divided into an output register and a garbage register; it is always assumed that the garbage register is traced out after the circuit is applied. The (possibly mixed) state which remains in the output register is called the output of the algorithm, and is denoted $\mathcal{Q}(\sigma)$.

We will sometimes also allow for algorithms which are allowed to mix probabilistic and quantum computation in a straightforward way: a classical probabilistic circuit first uses a string of classical randomness to decide which quantum circuit to run on the given quantum input state, and the chosen quantum circuit is then executed. We will call such algorithms probabilistic-quantum and refer to output states with or without specified randomness as above. The computational power of such algorithms can already be captured by quantum algorithms alone by reversibly implementing the classical pre-processing; the classical probabilistic mixtures are then absorbed into the density operator of the quantum state as it evolves under the quantum circuit. However, the distinction does have a difference: the final density operator outputted by the resulting quantum algorithm could exhibit some important property which is *not* true for any of the outputs of the original probabilistic-quantum algorithm. We will discuss an explicit example later.

An algorithm will be referred to as polynomial-time (or efficient) if all of the relevant circuit families are polynomial-time uniform [**?**]; this applies to all classes of algorithm discussed above.

3 Quantum encryption

In this section, we discuss a notion of encryption for quantum states with computational assumptions. Interestingly, this topic has not received significant attention as yet. In Section 3.1, we will recall how to construct a classical function which appears pseudorandom to quantum adversaries, by means of a function which is one-way against quantum adversaries. In Section 3.2, we define a notion of symmetric-key quantum encryption, together with associated notions of IND-CPA and IND-CCA1 security. We then describe a scheme which is IND-CCA1-secure under the assumption that quantum-secure one-way functions exist. While this particular scheme is new, encryption of quantum states with computational assumptions was also recently (and independently) considered by Broadbent and Jeffrey [7]. A complete framework, including considerations about semantic security, will appear in an upcoming work [3].

3.1 Quantum-secure pseudorandomness

We begin with two primitives for encryption: quantum-secure one-way functions, and quantum-secure pseudorandom functions. These are both classical, efficiently computable functions

[2] when there's more than one circuit for a given input size, there should be some efficient way to decide which inputs of that size are assigned to which circuit.

which are in some sense resistant to quantum analysis. In the case of one-way functions, we demand that inversion is hard; in the case of pseudorandom functions, we demand that distinguishing from perfectly random functions is hard.

Definition 1. *A PT-computable function $f : \{0,1\}^* \to \{0,1\}^*$ is a quantum-secure one-way function (qOWF) if for every QPT $\mathcal{A}$,*

$$\Pr_{x \in_R \{0,1\}^n} \left[\mathcal{A}(f(x), 1^n) \in f^{-1}(f(x)) \right] \leq \mathrm{negl}(n) \,,$$

where the probability is taken over $x \in_R \{0,1\}^n$ as well as the measurements of $\mathcal{A}$.

Definition 2. *A PT-computable function family $f_k : \{0,1\}^n \to \{0,1\}^m$ is a quantum-secure pseudorandom function (qPRF) if for every QPT $\mathcal{A}$,*

$$\left| \Pr_{k \in_R \{0,1\}^n} [\mathcal{A}^{f_k}(1^n) = 1] - \Pr_{g \in_R \mathcal{F}_{n,m}} [\mathcal{A}^{g}(1^n) = 1] \right| \leq \mathrm{negl}(n) \,,$$

where $\mathcal{F}_{n,m}$ denotes the space of all functions from $\{0,1\}^n$ to $\{0,1\}^m$.

Classically, one-way functions are the fundamental primitive underpinning encryption. A series of basic results shows that one-way functions can be turned into pseudorandom functions, which can then be used for defining probabilistic encryption schemes. This series of results carries over to the quantum-secure case without much of a change (although some proofs are somewhat more involved.) For example, it is known how to construct qPRFs from qOWFs.

Theorem 1. *If quantum-secure one-way functions exist, then so do quantum-secure pseudorandom functions.*

Proof. (Sketch.) It is folklore that the well-known Håstad et al. result that pseudorandom generators can be constructed from any one-way function [10] carries over to the quantum-secure case. Roughly speaking, the reasoning is that the reduction in the proof is done in a "black-box" way, i.e., only by feeding inputs into the adversary and then analyzing the resulting outputs. The quantum-secure case then simply involves replacing PPTs with QPTs in the appropriate places. Proving that the standard GGM construction [9] of PRFs from pseudorandom generators is still secure in the setting of quantum adversaries is more involved; this was established by Zhandry [13]. □

3.2 Symmetric-key encryption of quantum states

It is well-known how to encrypt quantum states with information-theoretic security, via the so-called quantum one-time pad. To encrypt a single-qubit state ρ, we choose two classical bits at random, use them to select a random Pauli matrix $P \in \{\mathbb{1}, X, Y, Z\}$, and perform $\rho \mapsto P\rho P^\dagger$. To encrypt an n-qubit quantum state ρ, we select $r \in_R \{0,1\}^{2n}$ and apply

$$\rho \longmapsto P_r \rho P_r^\dagger \,, \tag{3.1}$$

where P_r denotes the element of the n-qubit Pauli group indexed by r.

One disadvantage of the quantum one-time pad is that parties must share two bits of randomness for every qubit which they wish to transmit securely. In particular, one cannot securely exchange multiple messages with the same key. To address this issue, we must settle for computational security assumptions and use pseudorandomness to select r. A general encryption scheme for quantum states is then defined as follows.

Definition 3. *A symmetric-key quantum encryption scheme is a triple of QPTs:*

- *(key generation)* $\mathsf{KeyGen} : 1^n \longmapsto k \in \{0,1\}^n$;
- *(encryption)* $\mathsf{Enc}_k : \mathfrak{D}(\mathcal{H}_m) \longrightarrow \mathfrak{D}(\mathcal{H}_c)$;

- *(decryption)* $\mathsf{Dec}_k : \mathfrak{D}(\mathcal{H}_c) \longrightarrow \mathfrak{D}(\mathcal{H}_m)$;

where m and c are polynomial functions of n, and the QPTs satisfy $\|\mathsf{Dec}_k \circ \mathsf{Enc}_k - \mathbb{1}_m\|_\diamond \leq \mathrm{negl}(n)$ for all $k \in \mathbf{supp}\,\mathsf{KeyGen}(1^n)$.

Public-key quantum encryption schemes are defined in an analogous manner. The encryption schemes we will need must produce ciphertexts which are computationally indistinguishable. In some cases, the ciphertexts will need to remain indistinguishable even to adversaries which possess oracle access to the encryption algorithm (and sometimes also even the decryption algorithm.) This security notion is captured by the following definition.

Definition 4. *A symmetric-key quantum encryption scheme is IND-secure if for all QPTs $\mathcal{A}, \mathcal{A}'$,*

$$\left|\Pr[(\mathcal{A}' \circ \mathsf{Enc}_k \otimes \mathbb{1}_s \circ \mathcal{A}) \cdot 1^n = 1] - \Pr[(\mathcal{A}' \circ \Xi_{\mathsf{Enc}_k|0^m\rangle\langle 0^m|} \otimes \mathbb{1}_s \circ \mathcal{A}) \cdot 1^n = 1]\right| \leq \mathrm{negl}(n)\,,$$

where $\Xi_\sigma : \rho \mapsto \sigma$ is the "forgetful" map, and s is a polynomial function of n. If $\mathcal{A}$ and $\mathcal{A}'$ have oracle access to Enc_k, then we say that the scheme is IND-CPA secure. If in addition $\mathcal{A}'$ has oracle access to Dec_k, then we say that the scheme is IND-CCA1 secure.

The two QPTs $\mathcal{A}$ and $\mathcal{A}'$ together model the adversary. The definition above captures the idea of a certain "security game" between an adversary and a challenger. The game proceeds in steps: (i.) the key is selected and the adversary receives access to the appropriate oracles, (ii.) after some computation, the adversary transmits the first part of a bipartite state ρ_{ms} to a challenger, (iii.) the challenger either encrypts this or replaces it with the encryption of $|0^m\rangle\langle 0^m|$, and then returns the result to the adversary, and (iv.) the adversary must decide which choice the challenger made. The scheme is considered secure if the adversary can do no better than random guessing. As shown in [3], this definition is equivalent to a security notion called *semantic security*; roughly speaking, this notion captures the idea that anyone that tries to compute anything about a plaintext gains no advantage by possessing its encryption. In addition, Definition 4 is equivalent to several natural variants, where e.g., the challenger chooses to encrypt one of two messages provided by the adversary, or where the game is played over multiple rounds. The latter guarantees security of transmitting multiple ciphertexts produced via encryption with the same key.

We now show how to use qPRFs to construct simple symmetric-key quantum encryption schemes that satisfy all of the above security conditions.

Theorem 2. *If quantum-secure pseudorandom functions exist, then so do IND-CCA1-secure symmetric-key quantum encryption schemes.*

Proof. Let $\{f_k\}$ be a qPRF. For simplicity we assume that each f_k is a map from $\{0,1\}^n$ to $\{0,1\}^{2n}$. Recall that for $r \in \{0,1\}^{2n}$, P_r denotes the element of the n-qubit Pauli group indexed by r. Consider the following scheme:

- $\mathsf{KeyGen}(1^n)$: output $k \in_R \{0,1\}^n$;
- $\mathsf{Enc}_k(\rho)$: choose $r \in_R \{0,1\}^n$; output $|r\rangle\langle r| \otimes P_{f_k(r)} \rho P_{f_k(r)}^\dagger$;
- $\mathsf{Dec}_k(|r\rangle\langle r| \otimes \sigma)$: output $P_{f_k(r)}^\dagger \rho P_{f_k(r)}$.

In the decryption algorithm, we may assume that the first register is always measured prior to decrypting. Correctness of the scheme is straightforward to check: decrypting with the same key and randomness simply undoes the Pauli operation.

We now sketch the proof that the scheme is IND-CCA1 secure; a complete proof will appear in [**?**]. The key observation is that each query to the encryption oracle is no more useful than receiving a pair $(r, f_k(r))$ for $r \in_R \{0,1\}^{2n}$, and that each decryption oracle is no more useful than receiving a pair $(r, f_k(r))$ for a string r of the adversary's choice. Thus the adversary learns at most a polynomial number of values of f_k. Now, if f_k is a perfectly random function, then these

values are completely uncorrelated to the one used to encrypt the challenge. The scheme is thus secure simply by the information-theoretic security of the quantum one-time pad. On the other hand, if f_k is a function in a qPRF, Definition 2 guarantees oracle indistinguishability from perfectly random functions. It follows that, if $(\mathcal{A}, \mathcal{A}')$ can break the actual scheme, then by computational indistinguishability they would also break the perfect scheme, which is impossible. □

We emphasize that the above proof shows that, even in the case where the adversary chooses the randomness r used by the Enc_k and Dec_k oracles, the scheme remains secure. Of course, the randomness for the challenge encryption must still be selected by the challenger. Finally, by combining Theorem 1 and Theorem 2, we have the following.

Theorem 3. *If quantum-secure one-way functions exist, then so do IND-CCA1-secure symmetric-key quantum encryption schemes.*

4 Quantum black-box obfuscation

In this section, we discuss the virtual black-box framework for obfuscating quantum computations. We begin in Section 4.1 with a definition of black-box quantum obfuscator, motivated both by the classical analogue and an intuitive notion of what a "good obfuscator" should achieve. In Section 4.2, we outline several interesting cryptographic consequences that would follow from the existence of such an obfuscator. Finally, in Section 4.3, we prove a few impossibility results which restrict the range of possibilities for the existence of black-box quantum obfuscators. Interestingly, our results leave open some possibilities, which include (restricted versions) of the most interesting applications. Indeed, it is conceivable that quantum obfuscation could be significantly more powerful than its classical counterpart.

4.1 Definitions

Any reasonable notion of obfuscation involves giving the obfuscated circuit $\mathcal{O}(C)$ to an untrusted party. We accept as fundamental the idea that this obfuscated circuit should implement some particular, chosen functionality f_C, and that the object $\mathcal{O}(C)$ allows the untrusted party to execute that functionality. In the black-box formulation of obfuscation, we demand that this is effectively all that the untrusted party will ever be able to do. The rigorous formulation uses the simulation paradigm: anything which can be efficiently learned from the obfuscated circuit, should also be efficiently learnable simply by evaluating f_C some polynomial number of times. This "virtual black-box" notion was first formulated by Barak et al. [4], and proved impossible to satisfy generically in the classical case.

In the quantum case, there are several complications. First, we are considering the obfuscation of quantum functionalities. This implies that the end user (and hence also any adversary) should be in possession of a quantum computer, and likewise for the simulator. Second, it is conceivable that the obfuscation may not just be another quantum circuit, which is simply a classical state describing a quantum computation. The obfuscator might instead output a quantum state, which is then to be employed by the end user to execute the desired functionality in some well-specified manner. These considerations motivate the following definition.

Definition 5. *A **black-box quantum obfuscator** is a pair of QPTs $(\mathcal{J}, \mathcal{O})$ such that whenever C is a polynomial-size n-qubit quantum circuit, the output of $\mathcal{O}$ is an m-qubit state $\mathcal{O}(C)$ satisfying*

1. *(polynomial expansion) $m = poly(n)$;*
2. *(functional equivalence) $\left\| \mathcal{J}(\mathcal{O}(C) \otimes \rho) - U_C \rho U_C^\dagger \right\|_{\mathrm{tr}} \leq \mathrm{negl}(n)$ for all $\rho \in \mathfrak{D}(\mathcal{H}_n)$;*

3. *(virtual black-box) for every QPT $\mathcal{A}$ there exists a QPT $\mathcal{S}^{U_C}$ such that*

$$\left|\Pr\left[\mathcal{A}(\mathcal{O}(C))=1\right]-\Pr\left[\mathcal{S}^{U_C}\left(|0^n\rangle\right)=1\right]\right|\leq \mathrm{negl}(n)\,.$$

We remark that one could consider variants where the "interpreter" algorithm $\mathcal{J}$ is fixed once and for all, or where $\mathcal{O}(C)$ itself consists of both a quantum "advice state" and a circuit which the end user should execute on the advice state and the desired input. It is straightforward to show that all of these variants are equivalent, in the sense that a black-box quantum obfuscator of each variant exists if and only if the other variants exist. Since we are primarily concerned with possibility vs impossibility, we will stick with the formulation in Definition 5.

(Gorjan: Insert more careful version (with ensembles and distributions) of Definition 5 here.)

Finally, we point out that the no-cloning theorem opens up the possibility of *computationally unbounded adversaries*. In the classical case, such an adversary could simply execute the circuit on every input, and thus learn far more than is possible for a polynomial-time black-box simulator. Quantumly, however, a computationally unbounded adversary is restricted both by the no-cloning theorem and the limitations of measurement. The adversary may not be able to acquire multiple copies of the obfuscated state, and the single state may be partially (or completely) destroyed when measured. It is thus not *a priori* clear that an unbounded adversary could always outmatch a polynomial-time black-box simulator. The appropriate definition is a straightforward modification of Definition 5, where we replace the third condition with the following:

3. *(information-theoretic virtual black-box) for every quantum adversary $\mathcal{A}$ there exists a QPT $\mathcal{S}^{U_C}$ such that*

$$\left|\Pr\left[\mathcal{A}(\mathcal{O}(C))=1\right]-\Pr\left[\mathcal{S}^{U_C}\left(|0^n\rangle\right)=1\right]\right|\leq \mathrm{negl}(n)\,.$$

(Gorjan: Note that our two-circuit impossibility proof holds even for these kinds of obfuscators, for a simple reason: there's already a QPT adversary that no QPT simulator can beat.)

(Gorjan: Somewhere in here we need to mention that, when using obfuscated states, we will frequently write things like $\mathcal{O}(C)|\varphi\rangle$, which has the obvious meaning, but technically stands for appropriately using the interpreter (or the circuit given by the obfuscator), together with the advice state, as prescribed by the definition.)

(Gorjan: Do we want to discuss inefficient obfuscators? I guess we can show that inefficient perfect indistinguishability obfuscators exist... and that these are black-box for any circuits that *do* have black-box obfuscations...)

(Gorjan: Somewhere in here we need to mention that, when using obfuscated states, we will frequently write things like $\mathcal{O}(C)|\varphi\rangle$, which has the obvious meaning, but technically stands for appropriately using the interpreter (or the circuit given by the obfuscator), together with the advice state, as prescribed by the definition.)

4.2 Applications

In this section, we motivate the study of quantum black-box obfuscation by giving a few example applications. Many of these are motivated by known classical applications of classical black-box obfuscators. Although our impossibility results will put some restrictions on these applications, they remain interesting. In fact, some of the applications (such as quantum-secure one-way functions) will be used in the impossibility proofs themselves. We point out that, while most

of the applications below are written in terms of quantum functionality (e.g., encryption of quantum states), one can just as well consider the weaker case of classical functionality, in this case achieved via quantum means (e.g., via a quantum algorithm for obfuscation.)

4.2.1 Quantum-secure one-way functions

The first application shows that, if there exists a classical algorithm for obfuscating quantum computations, then quantum-secure one-way functions exist. By the results discussed in Section 3, this also implies the existence of quantum-secure pseudorandom generators, quantum-secure pseudorandom functions, and IND-CCA1-secure symmetric-key quantum encryption schemes.

Proposition 1. *If there exists a classical probabilistic algorithm which is a quantum black-box obfuscator, then quantum-secure one-way functions exist.*

Proof. The proof is essentially the same as that of Lemma 3.8 in [4]. For all $a \in \{0,1\}^n$ and $b \in \{0,1\}$, we define

$$U_{a,b} : |x, y\rangle \longmapsto \begin{cases} |a, y \oplus b\rangle & \text{if } x = a; \\ |x, y\rangle & \text{otherwise.} \end{cases}$$

Define a function $f : \{0,1\}^* \to \{0,1\}^*$ by $f(a,b,r) = \mathcal{O}_r(U_{a,b})$ where $\mathcal{O}$ is the obfuscator[3] as in the hypothesis, and $\mathcal{O}_r$ denotes the same algorithm, but with randomness coins initialized to r. Clearly, inverting f requires computing b from $\mathcal{O}_r(U_{a,b})$. Moreover, with only black-box access to $U_{a,b}$ (for uniformly random a, b) the probability of correctly outputing b in polynomial time is at most $1/2 + \mathrm{negl}(n)$. By the black-box property of $\mathcal{O}$, we then have

$$\begin{aligned} \Pr_{a,b}[A(f(a,b,r)) = b] &= \Pr_{a,b}[A(\mathcal{O}_r(a,b)) = b] \\ &\leq \Pr_{a,b}\left[S^{U_{a,b}}(1^n) = b\right] + \mathrm{negl}(n) \\ &\leq \frac{1}{2} + \mathrm{negl}(n)\,, \end{aligned}$$

which completes the proof. □

We remark that the above proof fails if the obfuscator is a quantum algorithm—even if its output is itself classical. The issue is that one-way functions must be deterministic; while one can turn a classical probabilistic algorithm into a deterministic one by making the coins part of the input, this is not possible quantumly. We leave the problem of constructing cryptographically useful primitives from a fully quantum obfuscator (or even just from a quantum encryption scheme) as an interesting open question.

4.2.2 CPA-secure private-key quantum encryption

Can we say anything about encryption of data if we know that *quantum* algorithms for quantum black-box obfuscation exist? While we do not know how to extract one-way functions, we can nonetheless produce useful encryption schemes, as follows.

Proposition 2. *If quantum black-box obfuscators exist, then so do IND-CPA-secure symmetric-key quantum encryption schemes.*

[3] For simplicity of notation, we omit $\mathcal{J}$ and assume that $f(a,b,r) = \mathcal{O}_r(U_{a,b})$ is in fact a classical circuit for $U_{a,b}$.

Proof. (Sketch.) Let $(\mathcal{O}, \mathcal{J})$ be a quantum black-box obfuscator. We consider an adaptation of the unitary operator $U_{a,b}$ defined above, but now with Pauli group action instead of XOR, and with two n-bit registers:

$$U'_{r,k} : |x, y\rangle \longmapsto \begin{cases} |x, P_r^\dagger y\rangle & \text{if } x = k; \\ |x, y\rangle & \text{otherwise,} \end{cases}$$

Now consider the following scheme for encrypting n-qubit quantum states.

- $\mathsf{KeyGen}(1^n)$: output $k \in_R \{0,1\}^n$;
- $\mathsf{Enc}_k(\rho)$: choose $r \in_R \{0,1\}^n$; output $P_r \rho P_r^\dagger \otimes \mathcal{O}(U_{r,k})$;
- $\mathsf{Dec}_k(\sigma \otimes \tau)$: output the second register of $\mathcal{J}(\tau \otimes |k\rangle\langle k| \otimes \sigma)$.

To check correctness, we apply the functionality-preserving property of the obfuscator. A decryption of a valid encryption with the same key yields

$$\begin{aligned} \mathsf{Dec}_k(\mathsf{Enc}_k(\rho)) &= \mathrm{Tr}_1 \left[\mathcal{J}(\mathcal{O}(U_{r,k}) \otimes |k\rangle\langle k| \otimes P_r \rho P_r^\dagger) \right] \\ &= \mathrm{Tr}_1 \left[U_{r,k}(|k\rangle\langle k| \otimes P_r \rho P_r^\dagger) U_{r,k}^\dagger \right] \\ &= \mathrm{Tr}_1 \left[|k\rangle\langle k| \otimes \rho \right] \\ &= \rho \,. \end{aligned}$$

as desired. IND-CPA security follows from the black-box property of the obfuscator, as follows. Let $\mathcal{A}$ be an adversary with access to the encryption oracle. Since the output of the encryption is a product state, $\mathcal{A}$ can be simulated by an adversary $\mathcal{S}$ that has only the first register of the ciphertext (i.e., $P_r \rho P_r^\dagger$) and black-box access to the unitary $U'_{r,k}$. It's then clear that $\mathcal{S}$ can only succeed in the challenge stage of Definition 4 by discovering the secret input for $U'_{r,k}$ or by guessing the response to the challenge. In any case, $\mathcal{S}$ (and hence also $\mathcal{A}$) succeeds with probability at most $1/2 + \mathrm{negl}(n)$. □

4.2.3 Public-key encryption from private-key encryption

As we now show, combining black-box obfuscation with one-way functions yields even stronger encryption functionality.

Proposition 3. *If quantum black-box obfuscators and quantum-secure one-way functions exist, then so do IND-CPA-secure public-key quantum encryption schemes.*

Proof. (Sketch.) Under the hypothesis, Theorem 3 implies the existence of IND-CCA1-secure symmetric-key encryption schemes for quantum states. Let $(\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$ be such a scheme; for concreteness, we may take the scheme described in Theorem 2. For $x \in \{0,1\}^n$, let $\mathsf{Enc}_{(x)}$ denote the encryption circuit for key x; this is the circuit that accepts two input registers (one for randomness, and one for the plaintext) and outputs the ciphertext. Now define a public-key encryption scheme $(\mathsf{KeyGen}', \mathsf{Enc}', \mathsf{Dec}')$ as follows.

- $\mathsf{KeyGen}'(1^n)$: output $sk := k \in_R \{0,1\}^n$ (secret key) and $pk := \mathcal{O}\left(\mathsf{Enc}_{(sk)}\right)$ (public key);
- $\mathsf{Enc}'_{pk}(\rho)$: choose $r \in_R \{0,1\}^n$; output $pk(|r\rangle\langle r| \otimes \rho)$;
- $\mathsf{Dec}'_{sk}(\sigma)$: output $\mathsf{Dec}_{sk}(\sigma)$.

The correctness of this scheme follows directly from the functionality-preserving property of $\mathcal{O}$ and the correctness of the private-key scheme. To prove IND-CPA security for the public-key scheme, we rely on the black-box property. It implies that any QPT adversary $\mathcal{A}$ with access to the public key can be simulated by a QPT $\mathcal{S}$ having only black-box access to $\mathsf{Enc}_{(sk)}$. The QPT $\mathcal{S}$, in turn, can be simulated by a QPT $\mathcal{S}'$ which has both decryption and encryption oracles for the

private-key scheme $(\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$. It may not be immediately obvious that the decryption oracle is necessary; this is the case because black-box access to $\mathsf{Enc}_{(sk)}$ enables $\mathcal{S}$ to select the randomness used for encryption, thus gaining the ability to evaluate pairs $(r, f_{sk}(r))$ where f is the qPRF from the private-key scheme.

Now we have that, if $\mathcal{A}$ can distinguish ciphertexts during the challenge, then so can $\mathcal{S}'$; since the ciphertexts themselves are the same for the public-key scheme and the private-key scheme, this contradicts the IND-CCA1 security of the private-key scheme. □

A few remarks are in order. First, in [**?**] it is shown that IND-CPA-secure public-key quantum encryption schemes exist under the assumption that quantum-secure trapdoor permutations exist. This is a stronger assumption than one-way functions. Proposition 3 can then be thought of as replacing this strengthening of assumptions with an obfuscator. In [7] it is shown how to use quantum-secure classical public-key encryption to produce quantum public-key encryption (by encrypting the key for the quantum one-time pad); this amounts to the same assumption on primitives as in [**?**]. An important difference between [**?** 7] and Proposition 3 is that the scheme from Proposition 3 may have public keys which are quantum states. Such schemes have not been considered before, and (due to no-cloning) would have significantly different features from their classical counterparts.

An interesting question is if there could be public-key encryption for classical data with classical ciphertexts, but where the encryption procedure is performed by a quantum algorithm. While this question remains open, our impossibility results will show that this cannot be achieved in a generic way via Proposition 3.

4.2.4 Quantum fully homomorphic encryption

We briefly recall the idea of fully homomorphic encryption (FHE). For thorough definitions and the appropriate notions of security in the fully quantum case, see [7]. Without considering all of the details, we will view QFHE as an encryption scheme (just as in Definition 3), but where KeyGen produces an extra "evaluation" key k_{eval}, and there is an "evaluation" algorithm:

- $\mathsf{Eval}_{k_{\text{eval}}} : \mathfrak{D}(\mathcal{H}_m \otimes \mathcal{H}_g) \longrightarrow \mathfrak{D}(\mathcal{H}_m)$.

We imagine a party (henceforth, *server*) in possession of k_{eval} and a ciphertext $\mathsf{Enc}_k(\rho)$ provided by another party (henceforth, *client*.) The evaluation algorithm then enables the server to produce the ciphertext $\mathsf{Enc}_k(G_k \rho G_k^\dagger)$, where G is a gate of the server's choice. A classical string describing the choice of gate G (and which qubits $k, k+1, \ldots$ of ρ it should be applied to) is input into the register $\mathcal{H}_g$. In general, we may consider the case where k_{eval} is itself a quantum state. Depending on the details of the scheme, this key may be partly or fully consumed by Eval; indeed, this is the case in [7]. Depending on the consumption rate, this might violate the (classically standard) *compactness* requirement for FHE, namely that the amount of communication between the client and the server should scale only with the size of the ciphertext, and not with the size of the computation the server wishes to perform.

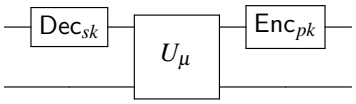
Proposition 4. *If quantum black-box obfuscators and one-way functions exist, then so do IND-CPA-secure quantum fully homomorphic encryption schemes.*

Proof. (Sketch.) We will consider the public-key case, which turns out to be simpler. Let $(\mathcal{O}, \mathcal{J})$ be a quantum obfuscator, and $(\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$ an IND-CPA-secure public-key scheme. We adapt KeyGen to produce an evaluation key, and describe the evaluation algorithm. We will require a universal circuit U_μ for performing gates on m-qubit states; this circuit accepts two inputs: an m-qubit state, and a description of a gate and indices of the qubits to which the gate should be applied. In our usage, m will be the number of qubits of the ciphertext state.

- $\mathsf{KeyGen}'(1^n)$: output $\mathsf{KeyGen}(1^n) = (sk, pk)$ and $k_{\text{eval}} = \mathcal{O}(\mathsf{Enc}_{pk} \circ U_\mu \circ \mathsf{Dec}_{sk})$;

- $\mathsf{Eval}_{k_{\text{eval}}} : \rho \otimes |G\rangle\langle G| \longmapsto \mathcal{J}(k_{\text{eval}} \otimes \rho \otimes |G\rangle\langle G|)$.

where $|G\rangle\langle G|$ is again just a classical string instructing U_μ to apply the desired gate. A circuit for $\mathsf{Enc}_{pk} \circ U_\mu \circ \mathsf{Dec}_{sk}$ is given below; the gate register is represented by the bottom wire.



We now want to show that $(\mathsf{KeyGen}', \mathsf{Enc}, \mathsf{Dec}, \mathsf{Eval})$ is a public-key QFHE scheme. The homomorphic property follows directly from the definition of Eval and the functionality-preserving property of the obfuscator. The security of the encryption scheme follows from IND-CPA security of $(\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$ and the black-box property of $(\mathcal{O}, \mathcal{J})$. The black-box property implies that each execution of the Eval algorithm is no more useful than providing the server with an encryption of $G\rho G^\dagger$. However, in the IND-CPA setting, the adversary can already use the CPA oracle to produce encryptions of *arbitrary* plaintexts of her choice (as opposed to just ones which are modifications of the plaintext provided by the client.) There is one additional wrinkle: by repeatedly applying gates (or even just the identity), the adversary can also produce multiple encryptions during the challenge round. However, as shown in [7], single-message IND-CPA is equivalent to multiple-message IND-CPA. By the assumption that $(\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$ is IND-CPA secure, it follows that the homomorphic scheme is also secure.

We remark that, in general, the encryption procedure Enc_{pk} may require an external source of randomness. This is certainly the case in classical encryption, but may not be required if the Enc algorithm is allowed to perform measurements. In any case, since we are starting with an IND-CPA public-key scheme, the adversary already has access to the public key and the ability to encrypt with randomness of her choice; the ability to choose randomness in Eval is of no additional benefit. □

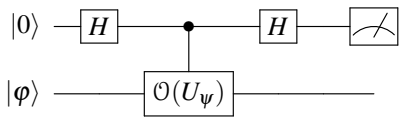
4.2.5 Public-key quantum money

Quantum money. The idea of "quantum money" first arose in work by Wiesner [12]. The core idea is simple: use a quantum state for representing currency in such a way that the no-cloning theorem of quantum mechanics prevents counterfeiting. These ideas were refined and developed further in several works [1, 2, 6, 8, 11]; some of these works also included explicit proposals based on various hardness assumptions.

Informally, a *quantum money scheme* consists of two algorithms: Mint, which produces quantum states, and Verify, which accepts an input state and then either accepts or rejects. If the different states produced by Mint are distinguishable, then we refer to them as *bills*; if they are indistinguishable, then we call them *tokens* (if Verify consumes them) or *coins* (if Verify does not consume them.) In all quantum money schemes, we imagine an authority (typically called the bank) which runs Mint repeatedly to produce money; in addition, the Verify algorithm should accept only on states produced by the bank. Depending on the particular scheme, this might only be true if Verify is executed by the bank (private-key money), or it might be true for any party (public-key money.)

In this language, Wiesner's original idea [12] was for a private-key scheme for bills, which is as follows. Each execution of Mint produces two random classical bitstrings $r, s \in \{0,1\}^{2n}$ as well as an n-qubit quantum state $|\psi_r\rangle$, with each qubit initialized in one of the states $|0\rangle, |1\rangle, |+\rangle, |-\rangle$, as determined by the bits of r. The bank records the pair (r, s) in a secret table, and publishes $(s, |\psi_r\rangle)$. The bank verifies by using s to look up the correct r in the table, and then performing the measurements in the correct basis and checking the results against r.

Public-key money from circuit obfuscation. While private-key money schemes are relatively straightforward to construct, public-key proposals appear to be much more difficult, and require computational assumptions. In analogy to its role in producing public-key encryption schemes from private-key ones (Proposition 2), an obfuscator can sometimes be used to turn private-key money schemes to public-key ones. The use of an obfuscator to create a particular quantum money scheme was considered by Mosca and Stebila [11]. Their scheme (in our language) is as follows. Each execution of Mint produces a Haar-random n-qubit quantum state $|\psi\rangle$, together with the obfuscation $\mathcal{O}(U_\psi)$ of a circuit[4] for $U_\psi = \mathbb{1} - 2|\psi\rangle\langle\psi|$. The bill consists of the pair $(\mathcal{O}(U_\psi), |\psi\rangle)$. Verify$(|\varphi\rangle)$ consists of executing the following:



and accepting iff the measurement returns 1. It's easy to check that the above succeeds only on valid states; moreover, in that case, the state $|\psi\rangle$ is output in the second register, so that verification can be repeated. To show resistance of the above scheme to counterfeiting, one can use Aaronson's Complexity-Theoretic No-Cloning Theorem [1], which states that cloning the state $|\psi\rangle$ while in possession of oracle access to $|U_\psi\rangle$ requires $\Omega(2^{n/2})$ queries. The first published proof of this theorem (as well as its first appearance in the form required here) was in [2].

Unfortunately, we will later show that obfuscation of quantum circuits in the form required by Mosca and Stebila is impossible. What remains possible is a setting in which both $|\psi\rangle$ and $\mathcal{O}(U_\psi)$ are quantum states, and another circuit (which is publicly known and independent of $|\psi\rangle$) is used for verification. Moreover, as we will also show, any black-box obfuscation scheme which outputs states that can be efficiently cloned is also impossible. We thus conjecture the following.

Conjecture 1. *If quantum black-box obfuscators exist, then so do public-key quantum money schemes.*

If the relevant obfuscation is a consumable state, then this would result in a token scheme. If it can be reused to perform verification repeatedly[5], then the result would be a bills scheme. We remark that, in any case, all of the public-key money states discussed above should be authenticated by the bank; otherwise a merchant would only know that he was handed *some* pair (state, circuit) where the circuit executed on the state outputs "accept"—a clearly inadequate state of affairs.

4.3 Impossibility results

4.3.1 Impossibility of two-circuit obfuscation

Barak et. al. [4] shows that black-box obfuscation is impossible by showing an explicit circuit family that cannot be black-box obfuscated. Here we present a similar result for **black-box quantum two-circuit obfuscation**, defined as in Definition 5 with the following strengthening of the virtual black-box condition:

[4]For most $|\psi\rangle$, the circuit U_ψ will not have polynomial length. However, as pointed out by [1], one can instead select $|\psi\rangle$ from an approximate t-design without a significant loss in security.

[5]This might seem to contradict no-cloning, but it does not: it is conceivable that the state can be used as an input to a unitary circuit where the desired output register contains a classical string with very high probability; this string can then be measured, copied and the unitary reversed to (approximately) recover the state.

3. *(two-circuit virtual black-box) for every pair of quantum circuits C_1 and C_2 and every quantum adversary $\mathcal{A}$ there exists a quantum simulator $\mathcal{S}^{U_{C_1},U_{C_2}}$ and a negligible ε_2 such that*

$$\left|\Pr[\mathcal{A}(\mathcal{O}(C_1)\otimes\mathcal{O}(C_2))=1]-\Pr\left[\mathcal{S}^{U_{C_1},U_{C_2}}\left(|0\rangle^{\otimes|C_1|+|C_2|}\right)=1\right]\right|\leq\varepsilon_2(n,\min\{|C_1|,|C_2|\})\,.$$

Theorem 4. *There exists an ensemble of distributions $\{\mathcal{H}_n\}_{n\in\mathbb{N}}$ over pairs quantum circuits, (C_n,D_n), of size poly(n), such that no pair of quantum circuits is a two-circuit black-box obfuscation of this ensemble of distributions.*

Proof. Let $(\mathcal{O},\mathcal{J})$ be a black-box quantum two-circuit obfuscator. We will show two-circuit impossibility for the following unitary operators. Here a and b are chosen uniformly at random from $\{0,1\}^n$. The registers indexed by x and y are of size n. The register indexed by C accepts a circuit description (under some fixed encoding), and needs to be able to handle inputs of size $|\mathcal{O}(C_{a,b})|$ where $C_{a,b}$ is a fixed, explicit poly(n)-size circuit for $U_{a,b}$. The second register of $V_{a,b}$ has size one.

$$U_{a,b}:|x,y\rangle\longmapsto\begin{cases}|x,y\oplus b\rangle & \text{if } x=a;\\ |x,y\rangle & \text{otherwise.}\end{cases}\tag{4.1}$$

$$V_{a,b}:|C,z\rangle\longmapsto\begin{cases}|C,z\oplus 1\rangle & \text{if } C(a)=b;\\ |C,z\rangle & \text{otherwise.}\end{cases}\tag{4.2}$$

Note that both of these unitaries can be implemented by efficient quantum circuits, $C_{a,b}$ and $D_{a,b}$, respectively, since the analogous classical function is efficiently computable. Further, define $Id_{2n}:|x\rangle|y\rangle\longmapsto|x\rangle|y\rangle$ to be the identity unitary on $2n$ qubits. Clearly, this can be implemented by an efficient quantum circuit, which we call Z_{2n}.

Now we notice that, for every QPT algorithm $\mathcal{S}$ there exists a negligible ε_1 so that:

$$\left|\Pr\left[\mathcal{S}^{U_{a,b},V_{a,b}}\left(|0\rangle^{\otimes|C_{a,b}|+|D_{a,b}|}\right)=1\right]-\Pr\left[\mathcal{S}^{Id_{2n},V_{a,b}}\left(|0\rangle^{\otimes|Z_{2n}|+|D_{a,b}|}\right)=1\right]\right|\leq\varepsilon_1(n,\min\{|C_{a,b}|,|D_{a,b}|\})\,.\tag{4.3}$$

Where the probability is taken over choice of a,b and the measurement outcome of the quantum algorithms. This is because with only polynomial queries, $\mathcal{S}$, which does not have knowledge of a or b, is forced to distinguish between unitaries which act identically on all but an exponentially small fraction of the total space. This is an easy corollary of the tightness of the Grover bound for unstructured quantum search [5].

However, consider the QPT algorithm $\mathcal{A}$ that, given as input the obfuscated states $\mathcal{O}(C)$ and $\mathcal{O}(D)$ simply runs $\mathcal{J}_{2n}(\mathcal{O}(D))$ on $|\mathcal{J}_{2n}(\mathcal{O}(C))\rangle|0\rangle$ and measures the second register, accepting iff it measures 1. Notice that this succeeds with constant probability $\alpha>0$ if given inputs $\mathcal{O}(C_{a,b})$ and $\mathcal{O}(D_{a,b})$, whereas this same algorithm A accepts with at most negligible probability when given input states $\mathcal{O}(D_{a,b})$ and $\mathcal{O}(Z_{2n})$, since the only way this happens is if $b=0^n$, which happens with negligible probability. Thus there exists a negligible function ε_2 so that:

$$\left|\Pr\left[\mathcal{A}(\mathcal{O}(D_{a,b}),\mathcal{O}(Z_{2n}))=1\right]-\Pr\left[\mathcal{A}\left(\mathcal{O}(D_{a,b})\otimes\mathcal{O}(C_{a,b})\right)=1\right]\right|\geq\alpha-\varepsilon_2(n,\min\{|C_{a,b}|,|D_{a,b}|\})\,.\tag{4.4}$$

Now consider the distribution $\mathcal{H}_n$ that is generated by choosing a,b uniformly at random from $\{0,1\}^n$, then outputting the respective pair of circuits $(C_{a,b},D_{a,b})$ with probability $1/2$ and probability $1/2$ outputting $(Z_{2n},D_{a,b})$. For this distribution, properties 4.3.2 and 4.4 together contradict the virtual black-box condition of the obfuscation procedure. □

4.3.2 Impossibility of obfuscation for cloneable outputs

Our goal in this section is to modify the black-box quantum two-circuit impossibility proof, in the prior section, to demonstrate impossibility of quantum circuit obfuscation with cloneable outputs:

Definition 6. *(Bill: This definition needs fixing!) A **black-box quantum obfuscator with cloneable outputs** is a pair of QPTs $(\mathcal{J}, \mathcal{O})$ such that whenever C is an n-qubit quantum circuit, the output is a $2m$-qubit state $\mathcal{O}(C) \otimes \mathcal{O}(C)$ satisfying*

1. *(polynomial slowdown) $m = poly(n, |C|)$;*
2. *(functional equivalence) $\left\| \mathcal{J}(\mathcal{O}(C) \otimes \cdot) - C \cdot C^\dagger \right\|_\diamond \leq \text{negl}(n, |C|)$;*
3. *(virtual black-box) for every QPT adversary $\mathcal{A}$ there exists a QPT simulator $\mathcal{S}^{U_C}$ such that*

$$\left| \Pr[\mathcal{A}(\mathcal{O}(C)) = 1] - \Pr\left[\mathcal{S}^{U_C}\left(|0\rangle^{\otimes |C|}\right) = 1\right] \right| \leq \text{negl}(n, |C|)\,.$$

Our main impossibility result is a modification of the proof in the prior section, following the classical proof [4].

Theorem 5. *There exists an ensemble of distributions $\{\mathcal{H}_n\}_{n\in\mathbb{N}}$ over quantum circuits, C_n, of size poly(n), such that no pair of quantum circuits is a black-box quantum obfuscation with cloneable outputs, of this ensemble of distributions.*

Proof. This proof works by carefully extending the two-circuit construction to show that a similar construction establishes impossibility for the cloneable case. First we give a general definition which will be useful:

Definition 7. *We define the* **combined quantum circuit** *of a finite collection of quantum circuits each with n input qubits, $\{C_1, C_2, ..., C_k\}$, to be the circuit that takes two registers, a control register of $\log k$ qubits, and an input register of n qubits, and controlled on the value of the first register applies the respective quantum circuit to the input register.*

Notice that if each circuit C_i in the collection is polynomial size, and k is bounded by a polynomial in n, then the associated combined quantum circuit is also polynomial sized.

Now consider the two unitaries $U_{a,b}$ and $V_{a,b}$ from Section 4.3.1, and their respective quantum circuits $C_{a,b}$ and $D_{a,b}$, as well as the circuit Z_{2n}, which simply implements the identity operator on $2n$ qubits. Also consider the combined circuits of $C_{a,b}$ and $D_{a,b}$ which we denote $C_{a,b}\#D_{a,b}$ and the combined quantum circuit of Z_{2n} and $D_{a,b}$ which we denote by $Z_{2n}\#D_{a,b}$. Using the reasoning of the argument from Section 4.3.1, these combined quantum circuits are indistinguishable from the perspective of any QPT simulator that is given only black-box access. On the other hand it is not immediately apparent that there exists an algorithm $\mathcal{A}$ that can distinguish inputs $\mathcal{O}(C_{a,b}\#D_{a,b})$ from $\mathcal{O}(Z_{2n}\#D_{a,b})$. This is because the naive algorithm that runs one copy of the output of the obfuscation on the other does not work, since the size of the obfuscated circuit generated by one copy of the obfuscation may be polynomially longer than the input register of the circuit generated by the second copy of the obfuscation.

To fix this issue, following the construction in the classical impossibility proof [4], our solution is to prove the following theorem about the existence of a distribution over circuits that allows a circuit of fixed input length to test whether a given circuit C of arbitrary polynomial size maps an input a to an output b. In particular, we show:

Lemma 6. *If quantum-secure one-way functions exists, then for each $n \in \mathbb{N}$ and $a, b \in \{0,1\}^n$ there exists a distribution $\mathcal{D}_{a,b}$ over circuits with the following properties:*

1. *Every $D \in$ **supp**$(\mathcal{D}_{a,b})$ is a circuit of size poly(n). Furthermore, there exists a QPT algorithm that, for every $n \in \mathbb{N}$ on input $a, b \in \{0,1\}^n$, samples the distribution $\mathcal{D}_{a,b}$.*

2. *There is a QPT algorithm $\mathcal{A}$ so that for all $n \in \mathbb{N}$, $a,b \in \{0,1\}^n$ and $D \in \textbf{\textit{supp}}(\mathcal{D}_{a,b})$, and for every circuit C, if $C|a\rangle|0^n\rangle = |a\rangle|b\rangle$, then $\mathcal{A}^D(C,1^n) = a$.*
3. *For any QPT S, $\Pr[S^{U_D}(1^n) = a] \leq neg(n)$, where the probability is over $a,b \in \{0,1\}^n$, $D \sim \mathcal{D}_{a,b}$, and the measurement of S.*

Proof. We follow closely the proof of Lemma 3.6 from the classical impossibility result [4], basically constructing a basic quantum private-key "homomorphic encryption" scheme. We think of each circuit $D \in \textbf{supp}(\mathcal{D}_{a,b})$ as the combined quantum circuit of the following three circuits which depend on a private key $K \in \{0,1\}^{2n}$ which will be used with the IND-CCA1-secure symmetric-key quantum encryption scheme from Theorem 2.

1. $E_{K,a}$ outputs $\mathsf{Enc}_K(|a\rangle)$.
2. $\mathsf{Hom}_K(C,\rho)$ takes a quantum circuit C, and a state ρ and outputs $\mathsf{Enc}_K(C(\mathsf{Dec}_K(\rho)))$ (Bill: TODO: add the randomness as input)
3. $B_{K,a,b}$ takes a quantum state ρ and outputs $|a\rangle$, if $\mathsf{Dec}_K(\rho) = |b\rangle$, and otherwise outputs $|0^n\rangle$.

Clearly given a and b, $\mathcal{D}_{a,b}$ can be sampled efficiently choosing K uniformly at random and outputting the combined quantum circuit $E_{K,a}\#\mathsf{Hom}_K\#B_{K,a,b}$, establishing Property 1 from the Lemma. Furthermore, notice that the QPT algorithm $\mathcal{A}$ that gets the description of a circuit C as input can check if $C(a) = b$ by using the three circuits comprising $D_{K,a,b}$ to simulate C gate-by-gate, using Hom_K initialized on the output of the $E_{K,a}$ circuit, and finally outputs the value of the circuit $B_{K,a,b}$, establishing Property 2.

It remains to verify Property 3, that no QPT simulator algorithm that has black-box access to each of the three algorithms comprising $D_{K,a,b}$ can discover a with non-negligible probability. We'll need the following lemma:

Lemma 7. *Let (Enc,Dec) be an IND-CCA1-secure symmetric-key quantum encryption, and* Hom *be as in the prior discussion. Then, for all n qubit quantum states ρ and every QPT algorithm $\mathcal{A}$:*

$$\left|\Pr[\mathcal{A}^{\mathsf{Hom}_K,\mathsf{Enc}_K}(\mathsf{Enc}_K(|0^n\rangle)) = 1] - \Pr\left[\mathcal{A}^{\mathsf{Hom}_K,\mathsf{Enc}_K}(\mathsf{Enc}_K(\rho)) = 1\right]\right| \leq \mathrm{negl}(n)\,.$$

Where the probabilities are over K chosen uniformly from $\{0,1\}^n$ and the measurement outcome of $\mathcal{A}$.

Proof. Assume there's an algorithm $\mathcal{A}$ that violates the claim. We'll show that this would break the IND-CCA1 security of the quantum encryption scheme.

To do this we first argue that we can replace the responses to all of $\mathcal{A}$'s queries to the Hom_K oracle with Encryptions of $|0^n\rangle$, with only a negligible loss in $\mathcal{A}$'s distinguishing gap. Consider the computation of $\mathcal{A}$ on input $\mathsf{Enc}_K(\rho)$ for each quantum state ρ on n qubits, and consider "hybrid" computations, where in the i-th hybrid, the first i queries of $\mathcal{A}$ to the Hom_K oracle are answered using the Hom_K oracle and the rest are answered using $\mathsf{Enc}_K(|0^n\rangle)$. Notice that any gap in distinguishing between the i and $i+1$st hybrid must be due to the $i+1$st query $\mathcal{A}$ makes to Hom_K, which is what differs between the hybrids. But we can now use this algorithm to create an adversary in violation of IND-CCA1 security of the encryption scheme. In particular, consider the algorithm that uses the Enc_k and Dec_k oracles to simulate all calls to the Hom_K oracle before receiving the challenge ciphertext, uses the challenge ciphertext as our answer to the $i+1$st query to Hom_K, and then answers all subsequent queries to Hom_K with $\mathsf{Enc}_K(|0^n\rangle)$. Thus any gap between the i and $i+1$st hybrid amounts to a distinguishing gap between quantum ciphertexts, in violation of IND-CCA1 security.

After this is established, we have that $\mathcal{A}$ can distinguish an encryption of $|0^n\rangle$ from an encryption of ρ, when given access to only an encryption oracle, again in violation of IND-CCA1. □

Notice that Lemma 7 suffices to establish Property 3, since giving the simulator algorithm black-box access to the three unitaries that comprise $D_{a,b}$ is equivalent to giving S black-box access to each circuit separately. Notice that black-box access to $E_{K,a}$ is no more powerful than giving it access to polynomially many queries of Enc_K, and giving black-box access to $B_{K,a,b}$ does not allow S to discover a with more than negligible probability, since it returns $|0^n\rangle$ on all but an exponentially small fraction of the space. Lemma 7 proves security in the presence of the Hom and Enc oracle. □

Now we are ready to adapt the two-circuit impossibility proof of Section 4.3.1 to the cloneable output case. First for given a, b let the distribution $\mathcal{D}_{a,b}$ be the distribution over circuits constructed in Lemma 6. Then consider the following two distributions over circuits:

1. $\mathcal{F}_n$: Choose a, b uniformly at random from $\{0,1\}^n$, sample a circuit D from $\mathcal{D}_{a,b}$ and output $C_{a,b}\#D_{a,b}$
2. $\mathcal{G}_n$: Choose a, b uniformly at random from $\{0,1\}^n$, sample a circuit D from $\mathcal{D}_{a,b}$ and output $Z_{2n}\#D_{a,b}$

By Property 2 of Lemma 6 there exists an algorithm $\mathcal{A}$ that, on input $\mathcal{O}(C) = \rho_{(1)} \otimes \rho_{(2)}$, runs $\mathcal{J}(\rho_{(1)})$ with the control register set to 1 and the input register containing $\mathcal{J}(\rho_{(2)})$, and accepts iff the first circuit in the combined circuit C outputs b on input a. Thus there exists a negligible function ε_1 so that:

$$\left|\Pr\left[\mathcal{A}(\mathcal{O}(\mathcal{F}_n)) = 1\right] - \Pr\left[\mathcal{A}\big(\mathcal{O}(\mathcal{G}_n)\big) = 1\right]\right| \geq \alpha - \varepsilon_1(n)\,.$$

While by Property 3 of Lemma 6, we know that for every QPT S there exists some negligible function ε_2 so that:

$$\left|\Pr\left[\mathcal{S}^{\mathcal{F}_n}\big(|0\rangle^{\otimes n}\big) = 1\right] - \Pr\left[\mathcal{S}^{\mathcal{G}_n}\big(|0\rangle^{\otimes n}\big) = 1\right]\right| \leq \varepsilon_2(n)\,.$$

□

5 Quantum indistinguishability obfuscation

5.1 Definitions

Definition 8. *An **indistinguishability quantum obfuscator** is a pair $(\mathcal{J}, \mathcal{O})$ where $\mathcal{J}$ is an interpreter and $\mathcal{O}$ is a quantum algorithm which on input an n-qubit quantum circuit C outputs an m-qubit quantum state $\mathcal{O}(C)$, such that*

1. *(polynomial slowdown) $m = poly(n, |C|)$.*
2. *(functional equivalence) there exists a negligible ε_1 such that $\left\|\mathcal{J}_n^{\mathcal{O}(C)} - U_C\right\|_\diamond \leq \varepsilon_1(n, |C|)$;*
3. *(indistinguishability) if a pair of circuits C_1 and C_2 satisfy $|C_1| = |C_2|$ and $\left\|U_{C_1} - U_{C_2}\right\|_\diamond \leq \varepsilon_3(n, |C|)$, then $\left\|\mathcal{O}(C_1) - \mathcal{O}(C_2)\right\|_{\mathrm{tr}} \leq \varepsilon_4(n, |C|)$.*

As before, we will select ε_3 and ε_4 appropriately later. For a definition of best-possible obfuscation, we replace condition (3) above with the following:

3. *(best-possible) for every pair of quantum circuits C_1 and C_2 that satisfy $|C_1| = |C_2|$ and $\left\|U_{C_1} - U_{C_2}\right\|_\diamond \leq \varepsilon_3(n, |C|)$ and every quantum adversary $\mathcal{A}$, there exists a quantum simulator $\mathcal{S}$ and a negligible ε_2 such that*

$$\left|\Pr[\mathcal{A}(\mathcal{O}(C_1)) = 1] - \Pr\left[\mathcal{S}\big(C_2\big) = 1\right]\right| \leq \varepsilon_2(n, |C|)\,.$$

The intuition behind the above definition is the following: any information $\mathcal{A}(\mathcal{O}(C_1))$ that is "leaked" by the obfuscation $\mathcal{O}(C_1)$ can actually be recovered from *any* functionally equivalent, similarly-sized circuit C_2. In this sense, among all such circuits, the circuit $\mathcal{O}(C_1)$ is one that leaks the least. It's not hard to see that an efficient obfuscator satisfies the best-possible condition if and only if it satisfies the indistinguishability condition. This justifies Definition 8 as a natural choice.

(Gorjan: To mention somewhere: GR07 observed that, if a circuit family *has* a black-box obfuscation, then a computational indistinguishability obfuscator must compute it. So it's conceivable that many of the interesting black-box applications carry over to the quantum case. Of course, one could say that this is exactly why the recent classical results have worked.)

5.2 Applications

Example: quantum witness encryption. The classical idea of witness encryption is from a paper by Sahai, Garg and others, and the idea of solving it with obfuscation is from the big paper by Sahai et al. In the quantum case, we set up the problem as follows. Suppose Alice wishes to encrypt a quantum plaintext $|x\rangle$, but not to a particular key or for a particular person; instead, the encryption is tied to a challenge question, and anyone that can answer the question correctly can decrypt the plaintext. Alice outputs a ciphertext $F_\phi|x\rangle$ where ϕ is a quantum 3-SAT formula, such that there exists an efficient algorithm Eval with the property that $\text{Eval}(F_\phi|x\rangle, |y\rangle) = |x\rangle$ if $|y\rangle$ is a satisfying assignment for ϕ. The security requirement is that if ϕ does not have a satisfying assignment, then the ensembles $F_\phi|x\rangle$ and $F_\phi|x'\rangle$ are quantum indistinguishable (formally, this now requires a definition of distinguishing *quantum* ensembles) whenever $|x\rangle$ and $|x'\rangle$ are quantum states on the same number of qubits. Note that the definition says nothing about the case where ϕ is satisfiable but a satisfying assignment is not known. While this may seem counterintuitive, Sahai and Garg etc. are nonetheless able to construct various interesting encryption schemes (like public-key encryption and identity encryption) from witness encryption.

The problem of quantum witness encryption can be solved using a quantum best-possible obfuscator $\mathcal{O}$, as follows. First, Alice selects a random Clifford (or Pauli) circuit C. She then writes down a quantum circuit M_C which accepts two registers (and some ancillas), such that $M|z\rangle|y\rangle|0\rangle = |C^{-1}z\rangle|y\rangle|0\rangle$ when $|y\rangle$ is a satisfying assignment for ϕ, and $M|z\rangle|y'\rangle|0\rangle = |z\rangle|y'\rangle|0\rangle$ for $|y'\rangle$ not a satisfying assignment for ϕ. The ciphertext $F_\phi|x\rangle$ will consist of the pair $(C|x\rangle, \mathcal{O}(M_C))$. A recipient with a satisfying assignment $|y\rangle$ can decrypt by computing $\mathcal{O}(M_C)|Cx\rangle|y\rangle|0\rangle$. On the other hand, if no satisfying assignment exists, then M_C acts like the identity operator on every input. By the definition of best-possible, a quantum adversary can learn nothing more from $\mathcal{O}(M_C)$ than she could from the trivial circuit with no gates. Moreover, by the design property of Cliffords (or Paulis) the adversary also observes $|Cx\rangle$ to be a maximally mixed state.

- Stephen has a description of how to build the circuit M_C, and that should be added.
- I guess the state $C|x\rangle$ and the circuit M_C are correlated. Is this a problem? This probably has to be addressed by defining quantum indistinguishability of quantum ensembles, and then showing that quantum indistinguishability of the classical ensemble $\mathcal{O}(M_C)$ plus 2-design property on $C|x\rangle$ implies quantum indistinguishability of the quantum ensemble $(C|x\rangle, \mathcal{O}(M_C))$.
- what does M_C do if you feed in a state that has a little bit of projection into a satisfying assignment? I guess that, unless the size of the projection is $1/\text{poly}$, it's still indistinguishable from identity...
- I have some ideas on why the above is exactly the right definition (e.g., weakening to ϕ being just a 3-SAT formula opens it up to being solved by classical obfuscation.)

5.3 Equivalence of indistinguishability and best-possible

(Gorjan: some old stuff below should be removed, but most still applies)

In what follows, for the sake of simplicity we omit the perfect, statistical, and classical variants of the definitions; one can arrive at these versions simply by replacing quantum indistinguishability of the relevant ensembles to one of the other notions. We will always be obfuscating quantum circuits, so when the word "quantum" appears in front of "obfuscator", this refers to the type of indistinguishability. We say that two uniform quantum circuit families $\mathcal{C}'$ and $\mathcal{C}''$ are equivalent if they consist of functionally equivalent circuits of the same size; more precisely, for every n, $|\mathcal{C}'_n| = |\mathcal{C}''_n| = 1$ and $|C'_n| = |C''_n|$ and $U_{C'_n} = U_{C''_n}$.

- the exact-same-length condition seems too strong, but it does appear in GR too, along with a later comment about how it can be removed. I guess some care is needed.

Definition 9. *A classical probabilistic algorithm $\mathcal{O}$ that takes as input a quantum circuit C and outputs another quantum circuit $\mathcal{O}(C)$ is a quantum* **best-possible obfuscator** *for the family $\mathcal{C}$ if it satisfies properties (1) and (2) from Definition 15, as well as the following property:*

3. *for any learner (uniform quantum circuit family) $\mathcal{L}$, there is a simulator (uniform quantum circuit family) $\mathcal{S}$ and a negligible ϕ such that, for all uniform equivalent subfamilies $\mathcal{C}', \mathcal{C}''$ of $\mathcal{C}$, the two ensembles $\mathcal{L}(\mathcal{O}(\mathcal{C}'))$ and $\mathcal{S}(\mathcal{C}'')$ are quantumly indistinguishable.*

(Gorjan: some old stuff below)

Definition 10. *A classical probabilistic algorithm $\mathcal{O}$ that takes as input a quantum circuit C and outputs another quantum circuit $\mathcal{O}(C)$ is a quantum* **indistinguishability obfuscator** *for the family $\mathcal{C}$ if it satisfies properties (1) and (2) from Definition 15, as well as the following property:*

3. *for all uniform equivalent subfamilies $\mathcal{C}', \mathcal{C}''$ of $\mathcal{C}$, the two ensembles $\mathcal{O}(\mathcal{C}')$ and $\mathcal{O}(\mathcal{C}'')$ are quantumly indistinguishable.*

- in all of the above, we could have considered obfuscating quantum states, or even using quantum algorithms to obfuscate classical descriptions of a quantum circuit. Why is this the "right" case (or at least an interesting one)?

With the definitions set up as above, many of the proofs of Goldwasser and Rothblum go through with little to no changes.

Proposition 1. *There exists an inefficient perfect indistinguishability obfuscator for all quantum circuits.*

Proof. The obfuscator just picks the lexicographically first circuit which implements the same unitary as the given circuit. Looping through lexicographically ordered circuits can be done in PSPACE, and equivalence-checking can be done in QMA $\subset$ QIP = PSPACE too. □

- what's the smallest class that one can do this in?

Proposition 2. *If $\mathcal{O}$ is a best-possible quantum obfuscator for a circuit family $\mathcal{C}$, then it is also a quantum indistinguishability obfuscator for $\mathcal{C}$.*

Proof. Let $\mathcal{C}'$ and $\mathcal{C}''$ be uniform equivalent subfamilies of $\mathcal{C}$, and let $\mathcal{L}$ be the trivial learner that simply implements the identity operator. By the best-possible property, there is a simulator $\mathcal{S}$ such that $\mathcal{S}(\mathcal{C}'')$ is quantum indistinguishable from $\mathcal{L}(\mathcal{O}(\mathcal{C}')) = \mathcal{O}(\mathcal{C}')$. By the same property, we also have that $\mathcal{S}(\mathcal{C}'')$ is quantum indistinguishable from $\mathcal{L}(\mathcal{O}(\mathcal{C}'')) = \mathcal{O}(\mathcal{C}'')$. By the transitivity property of indistinguishability, it follows that $\mathcal{O}(\mathcal{C}')$ is indistinguishable from $\mathcal{O}(\mathcal{C}'')$. □

Proposition 3. *If $\mathcal{O}$ is an efficient quantum indistinguishability obfuscator for a circuit family $\mathcal{C}$, then it is also an efficient quantum best-possible obfuscator for $\mathcal{C}$.*

Proof. Let $\mathcal{C}'$ and $\mathcal{C}''$ be equivalent subfamilies of $\mathcal{C}$, and let $\mathcal{L}$ be a (quantum) learner whose output on $\mathcal{C}'$ is the ensemble $\mathcal{L}(\mathcal{O}(\mathcal{C}'))$. We define a (quantum) simulator by setting $\mathcal{S} = \mathcal{L} \circ \mathcal{O}$; its output on $\mathcal{C}''$ is then the ensemble $\mathcal{L}(\mathcal{O}(\mathcal{C}''))$. Since the ensembles $\mathcal{O}(\mathcal{C}')$ and $\mathcal{O}(\mathcal{C}'')$ are quantum indistinguishable, so are their images under $\mathcal{L}$. □

5.4 Impossibility of statistical obfuscators

Recall the following computational problems and corresponding completeness results.

Definition 11. (a,b)-Identity Check.
Input: an n-qubit quantum circuit C.
Promise: $\min_\alpha \|U - e^{i\alpha} I\|$ *is less than a or greater than b.*
Output: YES in the former case and NO in the latter.

Theorem 8. *The problem (a,b)-Identity Check is coQMA-complete if $b - a \leq 1/\mathrm{poly}(n)$.*

Note: apparently Rosgen showed that distinguishing mixed state computations is QIP-complete. Does this mean that we could have even stronger impossibility results if we asked for obfuscators that could obfuscate quantum circuits that included measurements?

Given an m-qubit state ρ, let $\mathrm{Tr}_{(l,m)}[\rho]$ denote the result of tracing out qubits l through m. Nothing is traced out if $l > m$.

Definition 12. (a,b)-Quantum State Distinguishability
Input: m-qubit quantum circuits C_1 and C_2, positive integer $k \leq m$.
Promise: let $\rho_i = \mathrm{Tr}_{(k+1,m)}[C_i|0^m\rangle\langle 0^m|C_i^\dagger]$; then $\|\rho_0 - \rho_1\|_{\mathrm{tr}}$ is less than a or greater than b.
Output: YES in the former case and NO in the latter.

Theorem 9. *The problem (a,b)-Quantum State Distinguishability is QSZK-complete if $a < b^2$.*

We will in fact only need the containment part of the above theorem.

Theorem 10. *If there exists a polynomial-time indistinguishability quantum obfuscator, then coQMA is contained in QSZK.*

Proof. (parameters should be checked.) We will actually show coQMA $\subset \mathrm{BQP}^{\mathrm{QSZK}}$; since BQP is contained in QSZK, the result will follow. Let a and b satisfy $b - a \leq 1/\mathrm{poly}(n)$. We will solve (a,b)-Identity Check using a subroutine that solves (α,β)-quantum state distinguishability.

Let C be the input, i.e., a classical description of an n-qubit quantum circuit. Create an identity circuit D with an equal number of inputs as C, and of equal length to C. Let O_C be a circuit that initializes a register with the classical state $|C\rangle$ containing the classical description of C, and applies the circuit of $\mathcal{O}$ which corresponds to the input length $|C|$. Likewise, let O_D be be a circuit that initializes a register with the classical state $|D\rangle$ containing the classical description of D, and applies the circuit of $\mathcal{O}$ which corresponds to the input length $|D| = |C|$. Note that, after tracing out ancillas, the outputs of these circuits are given by

$$\mathrm{Tr}_{\mathrm{anc.}}\left[O_C|0\rangle\langle 0|O_C^\dagger\right] = \mathcal{O}(C) \qquad \text{and} \qquad \mathrm{Tr}_{\mathrm{anc.}}\left[O_D|0\rangle\langle 0|O_D^\dagger\right] = \mathcal{O}(D)\,.$$

Now apply the subroutine for solving quantum state distinguishability to the pair (O_C, O_D). If it says "close", we output YES; otherwise we output NO. Let's show that this has solved (a,b)-identity-check. Note that the states $\mathcal{O}(C)$ and $\mathcal{O}(D)$ must have the same number of qubits, and denote that number by m.

- **completeness.** In this case, the obfuscated states satisfy $\|\mathcal{O}(C) - \mathcal{O}(D)\|_{\mathrm{tr}} \leq \alpha$. By the definition of the induced trace norm, this implies that $\|\mathcal{J}^n_{\mathcal{O}(C)} - \mathcal{J}^n_{\mathcal{O}(D)}\|_\diamond \leq \alpha$. By functional equivalence for C and D and the triangle inequality, it follows that $\|U_C - U_D\|_\diamond = \|U_C - I\|_\diamond \leq \alpha$, as desired.

- **soundness.** In this case, the obfuscated states satisfy $\|\mathcal{O}(C) - \mathcal{O}(D)\|_{\mathrm{tr}} \geq \beta$. We claim that this implies $\|U_C - U_D\|_\diamond > b$. Suppose this is not the case, i.e., that these operators are in fact close; then by the indistinguishability property, it would follow that $\mathcal{O}(C)$ and $\mathcal{O}(D)$ are close as well, a contradiction.

The above amounts to a $\mathrm{BQP}^{\mathrm{QSZK}}$ protocol for a coQMA-hard problem, thus placing coQMA in QSZK. □

6 Discussion

Open questions:

- can you achieve single-copy vbb obfuscation with quantum states?
- can you achieve quantum circuit-to-circuit obfuscation under the comp. indistinguishability condition?
- what happens if we think about obfuscating measurements, or CPTP circuits?

References

[1] Scott Aaronson. Quantum copy-protection and quantum money. In *Computational Complexity, 2009. CCC'09. 24th Annual IEEE Conference on*, pages 229–242. IEEE, 2009.

[2] Scott Aaronson and Paul Christiano. Quantum money from hidden subspaces. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 41–60. ACM, 2012.

[3] Gorjan Alagic, Anne Broadbent, Bill Fefferman, Tommaso Gagliardoni, Christian Schaffner, and Michael StJules. Computational security for quantum encryption. *To appear.*

[4] Boaz Barak, Oded Goldreich, Russell Impagliazzo, Steven Rudich, Amit Sahai, Salil Vadhan, and Ke Yang. On the (im)possibility of obfuscating programs. *J. ACM*, 59(2):6:1–6:48, May 2012. ISSN 0004-5411. doi:10.1145/2160158.2160159. URL http://doi.acm.org/10.1145/2160158.2160159.

[5] Charles H. Bennett, Ethan Bernstein, Gilles Brassard, and Umesh V. Vazirani. Strengths and weaknesses of quantum computing. *SIAM J. Comput.*, 26(5):1510–1523, 1997. doi:10.1137/S0097539796300933. URL http://dx.doi.org/10.1137/S0097539796300933.

[6] CharlesH. Bennett, Gilles Brassard, Seth Breidbart, and Stephen Wiesner. Quantum cryptography, or unforgeable subway tokens. In David Chaum, RonaldL. Rivest, and AlanT. Sherman, editors, *Advances in Cryptology*, pages 267–275. Springer US, 1983. ISBN 978-1-4757-0604-8. doi:10.1007/978-1-4757-0602-4_26. URL http://dx.doi.org/10.1007/978-1-4757-0602-4_26.

[7] Anne Broadbent and Stacey Jeffery. Quantum homomorphic encryption for circuits of low T-gate complexity. *Crypto 2015 (to appear)*, December 2015.

[8] Edward Farhi, David Gosset, Avinatan Hassidim, Andrew Lutomirski, and Peter Shor. Quantum money from knots. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 276–289, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1115-1. doi:10.1145/2090236.2090260. URL http://doi.acm.org/10.1145/2090236.2090260.

[9] Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. *Journal of the ACM*, 33(4):792–807, 1986. ISSN 0004-5411. doi:http://doi.acm.org/10.1145/6490.6503.

[10] Johan Håstad, Russell Impagliazzo, Leonid A. Levin, and Michael Luby. A pseudorandom generator from any one-way function. *SIAM J. Comput.*, 28:1364–1396, March 1999. ISSN 0097-5397. doi:http://dx.doi.org/10.1137/S0097539793244708. URL http://dx.doi.org/10.1137/S0097539793244708.

[11] Michele Mosca and Douglas Stebila. Quantum coins. *Error-Correcting Codes, Finite Geometries and Cryptography*, 523:35–47, 2010.

[12] Stephen Wiesner. Conjugate coding. *ACM Sigact News*, 15(1):78–88, 1983.

[13] Mark Zhandry. How to Construct Quantum Random Functions. In *FOCS 2012*, pages 679–687. IEEE, 2012.

A Old VBB definitions

We now define the notion of an interpreter, which is simply a quantum algorithm equipped with some additional data.

Definition 13. *An **interpreter** is a polynomial-time uniform family of unitary quantum circuits $\mathcal{J} = \{J_{n,m}\}_{n,m\in\mathbb{N}}$, such that for every n and m, $J_{n,m}$ has an n-qubit register A, an m-qubit register B, and an ancilla C of size $\text{poly}(n,m)$. For every $n \in \mathbb{N}$ and every m-qubit state ρ, we define a superoperator on A by*

$$\mathcal{J}_n^\rho : \sigma \mapsto \text{Tr}_{BC}\left(J_{n,m}\left[\sigma \otimes \rho \otimes |0\rangle\langle 0|_C\right] J_{n,m}^\dagger\right).$$

In applications, we think of $\mathcal{J}$ as enabling an end-user to apply a superoperator to an input state σ with the help of the m-qubit "advice" state ρ, presumably provided by some other party. We say that the state ρ implements the operator $\mathcal{J}_n^\rho$. A simple example is given by universal circuits: in this case, ρ is a classical description of a quantum circuit for implementing $\mathcal{J}_n^\rho$, and $J_{n,m}$ consists of a universal sequence of gates which are applied to the first register and controlled by the second register. While this is not explicitly required by the definition, all advice states in this work will be efficiently preparable.

We now wish to consider obfuscated advice. Given a (potentially non-unitary) quantum circuit C, let U_C denote the superoperator implemented by C. We will frequently refer to "quantum adversaries" and "quantum simulators." Both will mean a polynomial-time quantum algorithm which accepts a quantum state as input (along with a polynomial-size initialized ancilla) and outputs a classical bit. We will sometimes make use of quantum simulators which have oracle access to some unitary operator; such a simulator will be denoted by, e.g., $\mathcal{S}^U$. The quantum circuits of such a simulator are allowed to make use of a "black box" gate which applies U. Each use of a black box counts towards the length of the circuit, which must remain polynomial in the input size.

Definition 14. *A **black-box quantum obfuscator** is a pair $(\mathcal{J}, \mathcal{O})$ where $\mathcal{J}$ is an interpreter and $\mathcal{O}$ is a probabilistic-quantum algorithm which, on input an n-qubit quantum circuit C, outputs an m-qubit quantum state $\mathcal{O}(C)$ satisfying*

1. *(polynomial slowdown) $m = \text{poly}(n, |C|)$;*
2. *(functional equivalence) there exists a negligible ε_1 such that $\left\|\mathcal{J}_n^{\mathcal{O}(C)} - U_C\right\|_\diamond \leq \varepsilon_1(n, |C|)$;*
3. *(virtual black-box) for every quantum adversary $\mathcal{A}$ there exists a quantum simulator $\mathcal{S}^{U_C}$ and a negligible ε_2 such that*

$$\left|\Pr[\mathcal{A}(\mathcal{O}(C)) = 1] - \Pr\left[\mathcal{S}^{U_C}\left(|0\rangle^{\otimes|C|}\right) = 1\right]\right| \leq \varepsilon_2(n, |C|).$$

We will select the functions $\varepsilon_1, \varepsilon_2$ later, to be the largest possible for which our impossibility proofs still work.

Note that allowing $\mathcal{O}$ to be probabilistic-quantum allows it to output different states $\mathcal{O}(C; r)$ depending on the choice of classical randomness r. Why is this different from just letting $\mathcal{O}$ be fully quantum? Suppose that we were to change the above definition to that effect, and consider the case where C is used to flip a single bit. The obfuscator could then, with equal probability, output either a state which always outputs 1 or a state which always outputs 0. Because we combined this choice into a density matrix, the functional equivalence condition would technically be satisfied. On the other hand, the actual output state of the algorithm would *never* satisfy functional equivalence – a frustrating situation for the end-user, to say the least.

B [OLD NOTES]

B.1 Preliminaries

Given a probability distribution X on a finite set S and an element $s \in S$, let $s \sim X$ denote the experiment of sampling s according to the distribution X. For example, $\Pr_{s\sim X}[s \in S']$ denotes the probability that a sample of X belongs to some subset $S' \subset S$. The total variation distance between two probability distributions X and Y taking values in S is defined by

$$|X - Y| = \frac{1}{2}\sum_{s\in S}\left|\Pr[X = s] - \Pr[Y = s]\right|.$$

If A and B are random variables with the same range, the notation $|A - B|$ will mean the total variation distance between the distributions of A and B.

We will often refer to circuits as deciding some problem, in the following sense. Let $\{C_n\}_{n\in\mathbb{N}}$ be a uniform family of classical probabilistic circuits. Fix $x \in \{0,1\}^n$, and let $C_n x$ denote the random variable determined by running C_n on the input x (with each remaining input bit set to either 0 or to the outcome of a uniformly random coinflip) and reading out the value of the first output bit. If the input x is selected according to some probability distribution A, then the acceptance probability is $\Pr_{x\sim A}[C_n x = 1]$. It is implicit that the probability is now taken over both the choice of x and the coins of C_n.

We can also view quantum circuits in this way. In the remainder of these notes, "quantum circuit" will always mean "unitary quantum circuit." Any measurements will be specified explicitly, and performed after the quantum circuit is applied. This is sufficient to describe arbitrary quantum computations (which in general may include many rounds of unitary operations, adapted measurements, and classical pre- and post-processing.) Set $\{C_n\}_{n\in\mathbb{N}}$ to be a uniform family of quantum circuits, and let $p(n)$ denote the number of qubits acted on by C_n. Given a quantum state $|\psi\rangle$ on n qubits, we apply the circuit C_n to the state $|\psi\rangle|0\rangle^{\otimes p(n)-n}$, and then measure the first qubit in the computational basis. This procedure can be described by a $\{0,1\}$-valued random variable, which we will denote by $M(C_n|\psi\rangle)$. Specifically, for $a \in \{0,1\}$,

$$\Pr[M(C_n|\psi\rangle) = a] = \sum_{x\in\{0,1\}^{p(n)}\,:\,x_1=a}\left|\langle x|C_n|\psi\rangle|0\rangle^{\otimes p(n)-n}\right|^2.$$

- it is worthwhile to discuss here why there is no "more powerful" way to use circuit families to solve decision problems (e.g., by appealing to PromiseBQP/PromiseBPP-hardness).

A *probability ensemble* D is a sequence $\{D_n\}_{n\in\mathbb{N}}$ of bitstring-valued random variables, such that for some polynomial ℓ, each D_n takes values in $\{0,1\}^{\ell(n)}$. We may sometimes need such ensembles to be polynomial-time constructible, meaning that one can sample from D via a uniform family of probabilistic circuits. Recall that a function $\phi : \mathbb{N} \to [0,\infty)$ is *negligible* if it is smaller than inverse-polynomial in n. We identify four distinct notions of indistinguishability of two probability ensembles A and B:

1. *perfectly indistinguishable*: $A_n = B_n$ for all sufficiently large n;
2. *statistically indistinguishable*: there exists a negligible function ϕ such that $|A_n - B_n| \leq \phi(n)$ for all sufficiently large n;
3. *quantumly indistinguishable*: there exists a negligible function ϕ such that, given any uniform family $\{C_n\}_{n\in\mathbb{N}}$ of quantum circuits, for all sufficiently large n we have

$$\left|\Pr_{x\sim A_n}[M(C_n|x\rangle) = 1] - \Pr_{x\sim B_n}[M(C_n|x\rangle) = 1]\right| \leq \phi(n).$$

4. *classically indistinguishable*: there exists a negligible function ϕ such that, given any uniform family $\{C_n\}_{n\in\mathbb{N}}$ of classical probabilistic circuits, for all sufficiently large n we have

$$\left| \Pr_{x\sim A_n}[C_n x = 1] - \Pr_{x\sim B_n}[C_n x = 1] \right| \leq \phi(n)\,;$$

The quantum case can also be expressed naturally in terms of density operators. Let us write $M(C_n\rho)$ for the random variable corresponding to the same decision experiment as before, but now starting with a density operator ρ. Given a probability distribution A on $\{0,1\}^n$, set

$$\rho_A = \sum_{x\in\{0,1\}^n} \Pr[A = x]|x\rangle\langle x|\,.$$

The random variable $M(C_n\rho_A)$ then exactly captures the outcome of selecting a string at random according to A, and then running the decision experiment corresponding to C_n. In other words,

$$\Pr_{x\sim A}[M(C_n|x\rangle) = a] = \Pr[M(C_n\rho_A) = a]\,.$$

Proposition 4. *Indistinguishability of probability ensembles satisfies*

$$\textit{perfect} \Rightarrow \textit{statistical} \Rightarrow \textit{quantum} \Rightarrow \textit{classical}\,.$$

Proof. Let A and B be probability ensembles. The first implication is immediate from the definition. For the second, recall the definition of trace norm $\|\rho\|_1 = \mathrm{Tr}\sqrt{\rho\rho^\dagger}$, and note that the trace distance between the two relevant density operators is

$$\|\rho_A - \rho_B\|_1 = 2|A - B|\,.$$

It's easy to check that the trace norm is unitarily invariant, so applying the same quantum circuit to both ρ_A and ρ_B does not affect the trace distance. The final measurement is just a projection to some subspace, and so the difference in the acceptance probabilities is bounded above by twice the trace distance. For the third implication, by standard arguments we can replace any classical circuit family that distinguishes two ensembles with a classical reversible circuit family that does the same. Reversible circuits are a special case of quantum circuits. □

- examples of why the implications are strict: (1) trivial, (2) large statistical difference but no quantum distinguisher (graph isomorphism?), and (3) quantum distinguisher but no classical distinguisher (factoring, or Bill's idea?).

By the triangle inequality, all four notions of indistinguishability are transitive, i.e. if A is indistinguishable from B and B is indistinguishable from C, then A is indistinguishable from C. All four notions of indistinguishability are also closed under applying polynomial-time operations to both ensembles; the exception is that classically indistinguishable ensembles may become classically distinguishable after an efficient quantum algorithm is applied.

B.2 Black-box Quantum circuit obfuscation

Given a (not necessarily uniform) family of circuits $\mathcal{C}$, let $\mathcal{C}_n$ denote the subset of $\mathcal{C}$ consisting of all circuits that act on exactly n qubits. If each $\mathcal{C}_n$ consists of one circuit only, then C_n will refer to that unique circuit, and the expression $\mathcal{C}|x\rangle$ will mean $C_n|x\rangle$ where n is the number of qubits of the state $|x\rangle$.

For a quantum circuit C, let U_C denote the unitary operator implemented by C. The notation S^C will stand for a quantum circuit S which, in addition to a universal set of quantum gates, can also make use of an additional black-box gate which implements U_C. The black-box gate can be used as many times as needed, although each use does count toward the total length of S^C.

We are now ready to define a few different notions of quantum circuit obfuscation. Our definitions closely follow the classical ones in Goldwasser and Rothblum.

Definition 15. *A classical probabilistic algorithm $\mathcal{O}$ that takes as input a quantum circuit C and outputs another quantum circuit $\mathcal{O}(C)$ is a quantum* **black-box obfuscator** *for the circuit family $\mathcal{C}$ if it satisfies:*

1. *preserving functionality: there is a negligible function ϕ such that for any n and any $C \in \mathcal{C}_n$,*

$$\Pr[U_C \neq U_{\mathcal{O}(C)}] \leq \phi(n).$$

2. *polynomial slowdown: there is a polynomial p such that for any $C \in \mathcal{C}$, $|\mathcal{O}(C)| \leq p(|C|)$.*
3. *virtual black-box: For any adversary (uniform quantum circuit family) $\mathcal{A}$, there is a simulator (uniform quantum circuit family) $\mathcal{S}$ and a negligible ϕ such that*

$$\left|\Pr[M(\mathcal{A}|\mathcal{O}(C)\rangle) = 1] - \Pr[M(\mathcal{S}^C|0\rangle) = 1]\right| \leq \phi(n)$$

for every n and every $C \in \mathcal{C}_n$.

- in GR there aren't four versions of the last property – just the computational one. Why?
- we may later wish to relax the functionality-preserving condition, so that two unitaries are considered functionally equivalent so long as (say) there is no polynomial-length proof of their inequality. This would affect later definitions too.
- are the classically not-black-box-obfuscatable functions also not quantum black-box obfuscatable?
- if not, are there other examples of not-quantum-black-box-obfuscatable functions? In order for these examples to be interesting, I guess they shouldn't be "learnable," i.e., you can't figure out exactly what they are with a polynomial number of black-box uses.
- is there an example family of quantum circuits which *is* black-box obfuscatable?

B.3 Best-possible

In what follows, for the sake of simplicity we omit the perfect, statistical, and classical variants of the definitions; one can arrive at these versions simply by replacing quantum indistinguishability of the relevant ensembles to one of the other notions. We will always be obfuscating quantum circuits, so when the word "quantum" appears in front of "obfuscator", this refers to the type of indistinguishability. We say that two uniform quantum circuit families $\mathcal{C}'$ and $\mathcal{C}''$ are equivalent if they consist of functionally equivalent circuits of the same size; more precisely, for every n, $|\mathcal{C}'_n| = |\mathcal{C}''_n| = 1$ and $|C'_n| = |C''_n|$ and $U_{C'_n} = U_{C''_n}$.

- the exact-same-length condition seems too strong, but it does appear in GR too, along with a later comment about how it can be removed. I guess some care is needed.

Definition 16. *A classical probabilistic algorithm $\mathcal{O}$ that takes as input a quantum circuit C and outputs another quantum circuit $\mathcal{O}(C)$ is a quantum* **best-possible obfuscator** *for the family $\mathcal{C}$ if it satisfies properties (1) and (2) from Definition 15, as well as the following property:*

3. *for any learner (uniform quantum circuit family) $\mathcal{L}$, there is a simulator (uniform quantum circuit family) $\mathcal{S}$ and a negligible ϕ such that, for all uniform equivalent subfamilies $\mathcal{C}', \mathcal{C}''$ of $\mathcal{C}$, the two ensembles $\mathcal{L}(\mathcal{O}(\mathcal{C}'))$ and $\mathcal{S}(\mathcal{C}'')$ are quantumly indistinguishable.*

Example: quantum witness encryption. The classical idea of witness encryption is from a paper by Sahai, Garg and others, and the idea of solving it with obfuscation is from the big paper by Sahai et al. In the quantum case, we set up the problem as follows. Suppose Alice wishes to encrypt a quantum plaintext $|x\rangle$, but not to a particular key or for a particular person; instead, the encryption is tied to a challenge question, and anyone that can answer the question correctly can decrypt the plaintext. Alice outputs a ciphertext $F_\phi|x\rangle$ where ϕ is a quantum 3-SAT formula, such that there exists an efficient algorithm Eval with the property

that $\mathrm{Eval}(F_\phi|x\rangle, |y\rangle) = |x\rangle$ if $|y\rangle$ is a satisfying assignment for ϕ. The security requirement is that if ϕ does not have a satisfying assignment, then the ensembles $F_\phi|x\rangle$ and $F_\phi|x'\rangle$ are quantum indistinguishable (formally, this now requires a definition of distinguishing *quantum* ensembles) whenever $|x\rangle$ and $|x'\rangle$ are quantum states on the same number of qubits. Note that the definition says nothing about the case where ϕ is satisfiable but a satisfying assignment is not known. While this may seem counterintuitive, Sahai and Garg etc. are nonetheless able to construct various interesting encryption schemes (like public-key encryption and identity encryption) from witness encryption.

The problem of quantum witness encryption can be solved using a quantum best-possible obfuscator $\mathcal{O}$, as follows. First, Alice selects a random Clifford (or Pauli) circuit C. She then writes down a quantum circuit M_C which accepts two registers (and some ancillas), such that $M|z\rangle|y\rangle|0\rangle = |C^{-1}z\rangle|y\rangle|0\rangle$ when $|y\rangle$ is a satisfying assignment for ϕ, and $M|z\rangle|y'\rangle|0\rangle = |z\rangle|y'\rangle|0\rangle$ for $|y'\rangle$ not a satisfying assignment for ϕ. The ciphertext $F_\phi|x\rangle$ will consist of the pair $(C|x\rangle, \mathcal{O}(M_C))$. A recipient with a satisfying assignment $|y\rangle$ can decrypt by computing $\mathcal{O}(M_C)|Cx\rangle|y\rangle|0\rangle$. On the other hand, if no satisfying assignment exists, then M_C acts like the identity operator on every input. By the definition of best-possible, a quantum adversary can learn nothing more from $\mathcal{O}(M_C)$ than she could from the trivial circuit with no gates. Moreover, by the design property of Cliffords (or Paulis) the adversary also observes $|Cx\rangle$ to be a maximally mixed state.

- Stephen has a description of how to build the circuit M_C, and that should be added.
- I guess the state $C|x\rangle$ and the circuit M_C are correlated. Is this a problem? This probably has to be addressed by defining quantum indistinguishability of quantum ensembles, and then showing that quantum indistinguishability of the classical ensemble $\mathcal{O}(M_C)$ plus 2-design property on $C|x\rangle$ implies quantum indistinguishability of the quantum ensemble $(C|x\rangle, \mathcal{O}(M_C))$.
- what does M_C do if you feed in a state that has a little bit of projection into a satisfying assignment? I guess that, unless the size of the projection is $1/\mathrm{poly}$, it's still indistinguishable from identity...
- I have some ideas on why the above is exactly the right definition (e.g., weakening to ϕ being just a 3-SAT formula opens it up to being solved by classical obfuscation.)

B.4 Indistinguishability

Definition 17. *A classical probabilistic algorithm $\mathcal{O}$ that takes as input a quantum circuit C and outputs another quantum circuit $\mathcal{O}(C)$ is a quantum* **indistinguishability obfuscator** *for the family $\mathcal{C}$ if it satisfies properties (1) and (2) from Definition 15, as well as the following property:*

3. *for all uniform equivalent subfamilies $\mathcal{C}', \mathcal{C}''$ of $\mathcal{C}$, the two ensembles $\mathcal{O}(\mathcal{C}')$ and $\mathcal{O}(\mathcal{C}'')$ are quantumly indistinguishable.*

- in all of the above, we could have considered obfuscating quantum states, or even using quantum algorithms to obfuscate classical descriptions of a quantum circuit. Why is this the "right" case (or at least an interesting one)?

B.5 Relationships between the definitions

With the definitions set up as above, many of the proofs of Goldwasser and Rothblum go through with little to no changes.

Proposition 5. *There exists an inefficient perfect indistinguishability obfuscator for all quantum circuits.*

Proof. The obfuscator just picks the lexicographically first circuit which implements the same unitary as the given circuit. Looping through lexicographically ordered circuits can be done in PSPACE, and equivalence-checking can be done in QMA $\subset$ QIP = PSPACE too. □

- what's the smallest class that one can do this in?

Proposition 6. *If $\mathcal{O}$ is a best-possible quantum obfuscator for a circuit family $\mathcal{C}$, then it is also a quantum indistinguishability obfuscator for $\mathcal{C}$.*

Proof. Let $\mathcal{C}'$ and $\mathcal{C}''$ be uniform equivalent subfamilies of $\mathcal{C}$, and let $\mathcal{L}$ be the trivial learner that simply implements the identity operator. By the best-possible property, there is a simulator $\mathcal{S}$ such that $\mathcal{S}(\mathcal{C}'')$ is quantum indistinguishable from $\mathcal{L}(\mathcal{O}(\mathcal{C}')) = \mathcal{O}(\mathcal{C}')$. By the same property, we also have that $\mathcal{S}(\mathcal{C}'')$ is quantum indistinguishable from $\mathcal{L}(\mathcal{O}(\mathcal{C}'')) = \mathcal{O}(\mathcal{C}'')$. By the transitivity property of indistinguishability, it follows that $\mathcal{O}(\mathcal{C}')$ is indistinguishable from $\mathcal{O}(\mathcal{C}'')$. □

Proposition 7. *If $\mathcal{O}$ is an efficient quantum indistinguishability obfuscator for a circuit family $\mathcal{C}$, then it is also an efficient quantum best-possible obfuscator for $\mathcal{C}$.*

Proof. Let $\mathcal{C}'$ and $\mathcal{C}''$ be equivalent subfamilies of $\mathcal{C}$, and let $\mathcal{L}$ be a (quantum) learner whose output on $\mathcal{C}'$ is the ensemble $\mathcal{L}(\mathcal{O}(\mathcal{C}'))$. We define a (quantum) simulator by setting $\mathcal{S} = \mathcal{L} \circ \mathcal{O}$; its output on $\mathcal{C}''$ is then the ensemble $\mathcal{L}(\mathcal{O}(\mathcal{C}''))$. Since the ensembles $\mathcal{O}(\mathcal{C}')$ and $\mathcal{O}(\mathcal{C}'')$ are quantum indistinguishable, so are their images under $\mathcal{L}$. □

B.6 Example: Clifford circuits

Recall that the single-qubit Pauli operators are defined by

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Each Pauli operator is self-adjoint and unitary. A few useful relations are

$$X^2 = Y^2 = Z^2 = I \qquad XY = -YX = iZ \qquad XZ = -ZX = -iY \qquad YZ = -ZY = iX\,.$$

From these relations, it's easy to see that the set of matrices αM where $\alpha \in \{\pm 1, \pm i\}$ and $M \in \{I, X, Y, Z\}$ forms a group under matrix multiplication. This group is generated by $\{X, Y, Z\}$ and $\{\pm 1, \pm i\}$. In the n-qubit case, we first set

$$X_j = I^{\otimes j-1} \otimes X \otimes I^{\otimes n-j}$$

and likewise for Y_j and Z_j. We define the n-qubit Pauli group $\mathcal{P}_n$ to be the group generated by $\{X_j, Y_j, Z_j : j = 1, \ldots, n\}$ and $\{\pm 1, \pm i\}$.

The Clifford group on n qubits is defined to be the normalizer of the Pauli group inside the unitary group, i.e.,

$$\mathcal{C}_n = \{U \in U(2^n) : UPU^\dagger \in \mathcal{P}_n \text{ for all } P \in \mathcal{P}_n\}\,.$$

By direct computation on the Pauli generators, it's easy to check that the following gates are elements of $\mathcal{C}_n$ for any $n \geq 2$:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad P = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}, \qquad CNOT = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

It is a theorem (see Gottesman's papers) that the above gates (when applied to arbitrary qubits or pairs of qubits) actually generate the entire Clifford group. A Clifford circuit is any circuit which

is made up of gates from the above gate set. It is well-known that Clifford circuit computations can be efficiently simulated by a classical computer, but that adding any gate outside the Clifford group yields a quantum-universal set. In spite of their lack of computational power, Clifford circuits are quite relevant in quantum information, e.g., in quantum error correction and quantum cryptography.

In this section, we show how to put any Clifford circuit into a unique normal form. Something like this is already discussed in Gottesman's PI lectures. Selinger also provides a unique normal form (as well as generators and relations for $\mathcal{C}_n$) but he uses a different gate set. The approach below also seems more natural, as it's closely related to how Cliffords are usually discussed in the QI literature.

For us, a "unique normal form" is a map f from Clifford circuits to Clifford circuits, such that *(i.)* C and $f(C)$ always implement the same unitary operator, and *(ii.)* whenever C_1 and C_2 are circuits which implement the same unitary operator, $f(C_1)$ and $f(C_2)$ are identical as circuits. We will sketch out how this can be done using a polynomial-time classical algorithm. By definition, this immediately gives an indistinguishability obfuscator for Clifford circuits.

Moreover, by a result of Richard Low, given a black box that implements a Clifford group element U, we can "learn" the action of U on the Pauli generators in polynomial time. As our algorithm will make clear, knowing the action of U on the generators suffices to produce the normal form. This means that any learner that has access to a normal-form Clifford circuit for U can be simulated by a learner with black-box access to U. This obfuscation scheme thus also satisfies the conditions of black-box obfuscation.

Unfortunately, this obfuscation is in some sense trivial; while it is true that the precise form of the initial circuit is not learnable from the obfuscated circuit, it is nonetheless easy to learn the full functionality.

We can map each element of the n-qubit Pauli group to a $2n$-bit string by ignoring the phase and setting

$$X_i \mapsto (\underbrace{0,\ldots,0}_{i-1},1,0,\ldots,0) \qquad \text{and} \qquad Z_i \mapsto (\underbrace{0,\ldots,0}_{n+i-1},1,0,\ldots,0)\,.$$

By checking the relations on the generating set, one sees that this map yields an isomorphism

$$f : \mathcal{P}_n/\{\pm 1, \pm I\} \to \mathbb{Z}_2^{2n}\,.$$

It's also easy to compute how the conjugation action of a Clifford gate on a Pauli generator affects the corresponding binary string. Since conjugation is linear, this is described by a matrix. For example,

$$H \mapsto \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad P \mapsto \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \qquad CNOT \mapsto \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In general, for any fixed n, applying any of the above gates to a particular qubit (or pair of qubits for CNOT) will correspond to some easily computable $2n \times 2n$ binary matrix. Given a Clifford circuit C, we can multiply the matrices corresponding to each gate in C to get a matrix $M(C)$. This matrix satisfies the property

$$M(C)f(P) = f(M(CPC^\dagger))$$

for every Pauli $P \in \mathcal{P}_n$. In fact, it is also the case that $M(C_1) = M(C_2)$ whenever C_1, C_2 are two Clifford circuits that implement the same element of the Clifford group. This follows from the isomorphism

$$\mathcal{C}_n' \cong \mathrm{Sp}(2n, \mathbb{F}_2)\,,$$

where $\mathcal{C}'_n$ denotes $\mathcal{C}_n$ modulo $\mathcal{P}_n$ and arbitrary phases, and $\mathrm{Sp}(2n, \mathbb{F}_2)$ denotes the group of $2n \times 2n$ symplectic matrices over $\mathbb{F}_2$. Why symplectic? Well, because Clifford elements preserve both commutation and anti-commutation of Pauli group elements, and whether two Pauli group elements commute or anti-commute is captured by a symplectic form of their corresponding binary strings:

$$PQ = (-1)^{\omega(f(P),f(Q))} QP$$

where

$$\omega(x,y) = (x_1, \ldots, x_n | y_{n+1}, \ldots, y_{2n}) + (y_1, \ldots, y_n | x_{n+1}, \ldots, x_{2n})$$

and $(a|b)$ denotes the dot product modulo 2.

It now remains to produce a unique Clifford circuit from $M(C)$, and append the right element of $\mathcal{P}_n$. The former is done through a row reduction procedure. The key observation is that row reduction operations correspond to left-multiplication by matrices corresponding to gates. Once we have row-reduced $M(C)$ to the identity, we then invert the sequence of gates we applied to output a circuit C'. We then know that

$$C^{-1}C' = P$$

for some $P \in \mathcal{P}_n$. By applying each gate of CC' to the Pauli generators, we can compute P and append its inverse to C'. This constitutes a unique circuit which is equivalent to C up to overall phases.

- the above is clearly just a sketch, which we can flesh out if we decide this is really important stuff.

Why is this uninteresting Note that any canonical form obfuscator is not, in general, a black-box obfuscator. A learner which is given the canonical form of a circuit can, in general, learn something that a learner with only black-box access cannot: namely, the canonical form itself! It's useful here to think about what such an obfuscator does on a family of circuits which are *already* in canonical form.

Now suppose all of the functions computed by the relevant class of circuits are black-box learnable, in the sense that there is an efficient algorithm which can use black-box access to a function f to output a description of any circuit (and hence also the canonical circuit) for computing f. Strictly speaking, the canonical-form obfuscator is now also a black-box obfuscator. But now again consider a uniform family of circuits which are already in canonical form. In this case, black-box access can be used to recover the entire original circuit perfectly. This should mean that, in an intuitive sense, obfuscation is completely impossible for this circuit family. This explains why our definitions (as well as the classical ones) are meaningless when we talk about efficiently learnable functions.