2018 IEEE 59th Annual Symposium on Foundations of Computer Science

Low-degree testing for quantum states, and a quantum entangled games PCP for QMA

Anand Natarajan
Center for Theoretical Physics
MIT
Cambridge, USA
Email: anandn@mit.edu

Thomas Vidick
Department of Computing and Mathematical Sciences
California Institute of Technology
Pasadena, USA
Email: vidick@cms.caltech.edu.

***Abstract*—We show that given an explicit description of a multiplayer game, with a classical verifier and a constant number of players, it is QMA-hard, under randomized reductions, to distinguish between the cases when the players have a strategy using entanglement that succeeds with probability 1 in the game, or when no such strategy succeeds with probability larger than $\frac{1}{2}$. This proves the "games quantum PCP conjecture" of Fitzsimons and the second author (ITCS'15), albeit under randomized reductions.**

The core component in our reduction is a construction of a family of two-player games for testing n-qubit maximally entangled states. For any integer $n \geq 2$, we give such a game in which questions from the verifier are $O(\log n)$ bits long, and answers are $\mathrm{poly}(\log\log n)$ bits long. We show that for any constant $\varepsilon \geq 0$, any strategy that succeeds with probability at least $1-\varepsilon$ in the test must use a state that is within distance $\delta(\varepsilon) = O(\varepsilon^c)$ from a state that is locally equivalent to a maximally entangled state on n qubits, for some universal constant $c > 0$. The construction is based on the classical plane-vs-point test for multivariate low-degree polynomials of Raz and Safra (STOC'97). We extend the classical test to the quantum regime by executing independent copies of the test in the generalized Pauli X and Z bases over $\mathbb{F}_q$, where q is a sufficiently large prime power, and combine the two through a test for the Pauli twisted commutation relations.

Our main complexity-theoretic result is obtained by combining this family of games with techniques from the classical PCP literature. More specifically, we use constructions of PCPs of proximity introduced by Ben-Sasson et al. (CCC'05), and crucially rely on a linear property of such PCPs. Another consequence of our results is a deterministic reduction from the games quantum PCP conjecture to a suitable formulation of the constraint satisfaction quantum PCP conjecture.

I. Introduction

The PCP theorem [1], [2] makes a remarkable statement: any language that admits efficiently verifiable proofs of membership, i.e. any problem in NP, also admits proofs that can be verified by reading only a *constant* number of bits of the proof. Do similar encodings exist for problems that admit *quantum* proofs? Consider the local Hamiltonian problem. Is there a way to encode a witness for the minimal energy of a Hamiltonian in a way that the energy can be estimated to within inverse polynomial accuracy while accessing only a constant number of bits, or qubits, from the witness? The pursuit of this question, which, broadly speaking, asks for quantum extensions of the PCP theorem, has been one of the most fruitful and challenging problems animating quantum complexity theory in the past decade: it ties in to the theory of quantum error-correcting codes, has applications to quantum cryptography, and promises insights into the study of entanglement in ground states of local Hamiltonians [3].

The question can be formalized in multiple ways. A first formulation, the "constraint satisfaction" variant of the quantum PCP (QPCP) conjecture [4], asks for the complexity of constant-factor approximations to the minimal energy of a local Hamiltonian H, normalized so that $\|H\| = 1$. Despite considerable attention progress on the conjecture has been difficult [5], [6], [7].

More recently a second formulation has been put forward. The "multiplayer games" variant of the QPCP conjecture, introduced in [8], asks for the complexity of estimating, to within constant accuracy, the maximum success probability of provers (we use the terminology "provers" and "players" interchangeably) sharing entanglement in a multiplayer game, a quantity referred to as the *entangled value* of the game. The conjecture is a natural analogue of the "oracularized" formulation of the PCP theorem, which states that the maximum success probability of *classical* provers in a multiplayer game is NP-hard to approximate to within constant factors. (This can be thought of as a "scaled down" formulation of the equality $\mathrm{MIP} = \mathrm{NEXP}$ [9].)

In [10], building on [11] it was shown that the approximation problem for the entangled value of a multiplayer game remains NP-hard, provided there are at least three provers. This was extended to games with two provers only in [12] (this result will be used as a building block in the present paper). In [8], [13] it was shown that inverse-polynomial approximations are QMA-hard (provided there are at least five provers), a result that is akin to a "quantum Cook-Levin theorem for entangled games." These results motivate the following conjecture, first made in [8]:

Conjecture I.1 (Games QPCP conjecture (informal))**.** *Sup-*

2575-8454/18/$31.00 ©2018 IEEE
DOI 10.1109/FOCS.2018.00075

pose given as input an explicit description of a classical multiplayer game. Then it is QMA*-hard to determine whether provers sharing quantum entanglement (of arbitrary dimension) have optimal success probability at least $\frac{2}{3}$ or at most $\frac{1}{3}$ in the game.*

We show that the conjecture holds, under randomized reductions.

Theorem I.2 (Games QPCP under randomized reductions)**.** *Suppose given as input an explicit description of a classical multiplayer game. Then it is* QMA*-hard, under randomized reductions, to determine whether provers sharing quantum entanglement (of arbitrary dimension) have optimal success probability at least* 1 *or at most $\frac{1}{2}$ in the game.*

Theorem I.2 is stated and proved as Corollary 6.14 in the full version of this paper [14]. The choice of constant $\frac{1}{2}$ in Theorem I.2 is arbitrary, as for the kind of games we consider soundness amplification can be performed efficiently in parallel [15].

We explain the need for a randomized reduction. Informally, the reason is that we do not know of a strong enough QMA-hardness result for the local Hamiltonian problem to initiate our reduction. In fact, we give two alternate formulations of Theorem I.2 that would also establish the same QMA-hardness result, under deterministic reductions, provided that either:

(i) it is QMA-hard to approximate the minimum energy of a local Hamiltonian in Y-free form (Definition 6.8 in the full version [14]) to within constant accuracy (this is a variant of the quantum PCP conjecture for local Hamiltonians), or

(ii) it is QMA hard to approximate the ground energy of (not necessarily local) frustration-free Hamiltonian whose every term is a tensor product of generalized Pauli τ_X or τ_Z observables.

Note that point (i) amounts to a deterministic reduction from Conjecture I.2 to the constraint satisfaction quantum PCP conjecture, and establishes the first proven relation between the two conjectures (see [16] for an incomparable result that relates stronger variants of both conjectures). Point (ii) is arguably a weaker assumption, as the gap is not required to be a constant and the terms of the Hamiltonian are not required to be local. However, due to the restriction that the Hamiltonian is frustration-free, it is currently not known whether the problem is QMA-hard (or even QMA_1-hard — though the frustration-free assumption can be relaxed to having exponentially small ground state energy).

Our results build on two main tools: a framework for protocols to test ground states, introduced in [8] and further developed in [13], [17], and a new proof of soundness of the classical low-degree test of Raz and Safra against two entangled provers [12]. The main result that underlies the complexity-theoretic applications is a two-prover test for n-qudit maximally entangled states, where each qudit has dimension $q = p^t = \text{poly}\log(n)$ for a prime p and integer t, that has inverse robustness independent of n (for all ε that are at least inverse polylogarithmic in n) and in which the verifier sends only $O(\log(n))$ bits to the provers, who reply with $O(\log\log n)$ bits each (Theorem III.2). This is an exponential improvement over all previous results, and provides the first robust entanglement test with sub-linear communication. While the ability to "test" structured objects with sub-linear efficiency has become customary in classical computer science, we find it remarkable that the framework for such tests may be extended to test such a complex object as quantum entanglement.

We first describe this test in more detail, before expanding on the complexity-theoretic consequences.

Efficient, robust entanglement tests: The driving question behind our work is the following: "Is it possible to verify a quantum state using an amount of resources that scales sub-linearly in the number of qubits of the state?" We start with the "simplest" such state—the maximally entangled state. Results in self-testing have yielded increasingly efficient and robust tests for this state and other, more general families of highly entangled states. Here we loosely refer to the "efficiency" of a test as a measure of the total number of bits of communication involved in an execution of the test. The "robustness" of the test indicates how tightly success in the test characterizes the desired state: a test is $\delta(\varepsilon)$-robust if for all $\varepsilon \geq 0$, any strategy for the provers that succeeds with probability at least $1-\varepsilon$ in the test must use an entangled state that is within distance $\delta(\varepsilon)$ from the tested state (see Definition II.3). Using these measures, the best prior self-tests for a maximally entangled state of n qubits are a test with communication $O(\log n)$ and robustness $O(n^{5/2}\sqrt{\varepsilon})$ [18] and a test with communication $O(n)$ and robustness $O(\sqrt{\varepsilon})$ [17]. Other recent results in this direction include [19], [20], [21], [22].

Our test is the first to combine robustness $\delta(\varepsilon) = \text{poly}(\varepsilon)$ that is independent of n, and logarithmic communication. Achieving both simultaneously is crucial to applications: constant (in n) robustness allows us to achieve gap-preserving reductions; logarithmic communication allows us to achieve efficient reductions.

As in previous results, the test is designed to constrain successful provers to use observables satisfying suitable relations; a statement about the entangled state follows by using that the state is stabilized by (a subset of) these observables. In the case of the maximally entangled state, the observables are all n-fold tensor products of Pauli observables. For reasons to be discussed below we test for qudits of dimension $q = p^t$ a prime power of order $q = \text{poly}\log(n)$. This leads us to consider tensor products of single-qudit Pauli observables defined over the prime power field $\mathbb{F}_q$, which we denote using the symbol τ:

$$\tau_X(a) = \sum_{j\in\mathbb{F}_q} |j+a\rangle\langle j| \qquad \text{and} \qquad \tau_Z(b) = \sum_{j\in\mathbb{F}_q} \omega^{\mathrm{tr}(bj)}|j\rangle\langle j| \,, \tag{1}$$

where $a,b \in \mathbb{F}_q$, $\omega = e^{\frac{2i\pi}{p}}$, addition and multiplication are over $\mathbb{F}_q$, and $\mathrm{tr}(\cdot)$ denotes the trace of $\mathbb{F}_q$ over $\mathbb{F}_p$. The main difficulty we face is that there are $2\cdot q^n$ such observables, $\tau_X(a) = \tau_X(a_1)\otimes\cdots\otimes\tau_X(a_n)$ and $\tau_Z(b) = \tau_Z(b_1)\otimes\cdots\otimes\tau_Z(b_n)$ for $a,b\in\mathbb{F}_q^n$, an exponentially larger number than any test with polylogarithmic communication gives us direct access to. It is then natural to consider a test that certifies observables $\tau_X(a)$ and $\tau_Z(b)$ for $a,b\in T\subseteq\mathbb{F}_q^n$, where $|T| = \mathrm{poly}(n)$, and attempt to construct observables for all $a,b\in\mathbb{F}_q^n$ in an inductive fashion, as is done in e.g. [18], where T is the set of all strings of Hamming weight at most 2. Unfortunately, any naïve procedure will induce an error accumulation at each step of the induction, eventually resulting in a robustness parameter that depends polynomially on n (as is the case in [18]).

It is thus crucial to choose the set T carefully — informally, it seems natural to require that this set behave in a "pseudorandom" way. We take direct inspiration from the classical proof of the PCP theorem, and use a set T specified as the set of all codewords of a suitably chosen Reed-Muller code; this is the reason for using a sufficiently large qudit dimension q. Our proof eventually reduces the analysis to the soundness of the entangled-prover classical low-degree test [12]. We explain the test, and its analysis, in more detail in Section I-A below.

Testing ground states and a "gap preserving" reduction: We sketch how our test for entanglement is applied to obtain results on the complexity of multiplayer entangled games. In the classical case, the proof that the value of a multiplayer game is at least as hard to approximate as the maximum fraction of constraints simultaneously satisfiable in a local constraint satisfaction problem proceeds via the technique of oracularization: the verifier selects a constraint at random and asks one prover for an assignment to all variables in the constraint and the other for an assignment to a single one of the variables. Given the provers' answers, the verifier checks the natural satisfaction and consistency constraints. In the quantum case the analogous idea would require each prover to hold a copy of the ground state of a QMA-complete local Hamiltonian, and return qubits as requested by the verifier. This reduction does not work: it is not possible in general to check for "consistency" between the same qubit taken from two copies of an entangled state. In [8] the idea was introduced of encoding the ground state using an error-correcting code and distributing a share to each prover. Subsequent work [13] showed that this idea can be used to show QMA-hardness of inverse-polynomial approximations to the entangled value of a multiplayer game. Unfortunately the reduction in [13] is not "gap-preserving": a large promised energy gap in the starting instance of the local Hamiltonian problem does not lead to a large completeness-soundness gap in the resulting game. As a result, even assuming the "constraint satisfaction" QPCP does not lead to hardness for approximation factors larger than a fixed inverse polynomial. In [17] we leveraged an entanglement test with constant robustness to achieve a gap-preserving reduction; unfortunately communication in the test is linear, resulting in a game with exponential size, so that no new complexity-theoretic consequence is obtained.

Armed with an exponentially more efficient entanglement test we are able to provide a much more effective reduction, yielding games of polynomial size from instances of the local Hamiltonian problem. The reduction follows similar lines as previous work, but with a new difficulty. Our entanglement test only certifies a specific family of observables: tensor products of generalized Pauli observables (1) over $\mathbb{F}_q$, for q a sufficiently large prime power. This requires us to initiate any direct reduction with a specific class of Hamiltonians, in so-called Y-free form (see Definition 6.8 of [14]); informally, these are local Hamiltonians such that each local term is a tensor product of generalized τ_X and τ_Z observables. In the absence of general gap-preserving reductions between different variants of the local Hamiltonian problem (perturbation techniques [23] do not generally preserve the promise gap) we obtain a reduction to the hardness of constant-factor approximations to the ground energy of local Hamiltonian of this form only. Nevertheless, even though the entanglement test requires a qudit dimension that scales (poly-logarithmically) with n, we show that any qubit Hamiltonian in Y-free form can be embedded in a Hamiltonian in Y-free form over qudits of dimension 2^t for any $t\geq 1$. As a result, we immediately obtain point (i) discussed earlier: that Conjecture I.2 would follow from QMA-hardness of constant-factor approximations to local Hamiltonian whose every local term is a tensor product of τ_X and τ_Z Pauli observables (signed weights of up to polylogarithmic size are allowed).

Composition and PCP: To obtain strong results we develop more elaborate reductions, with the aim of removing the assumption on *locality* of the Hamiltonian whose ground state energy is being tested. As our entanglement test has direct access only to local Pauli observables, it cannot be used to evaluate the expectation value of nonlocal observables (acting on more than a constant number of qudits). We get around this as follows. Say the verifier would like to estimate the expectation value of a nonlocal tensor product observable such as $\tau_X(b)$, for some $b\in\mathbb{F}_q^n$. The verifier asks each prover to measure all its qudits in the X basis, obtaining an outcome $a\in\mathbb{F}_q^n$, and report the value of the inner product $c = b\cdot a$. This provides the verifier with an estimate of the energy of $\tau_X(b)$. However, it remains to ensure that the outcome reported by the prover was obtained honestly, i.e. by measuring all qudits on which the observable acts, without having the ability to “read” all the single-qubit outcomes obtained. This sounds very similar to the kind of NP statements that PCPs are designed to allow efficient verification of, and indeed we employ classical PCP techniques, more specifically the notion of *PCP of proximity* (PCPP).

In order to verify that a prover honestly computed the inner product $c = b \cdot a$, the verifier asks it to provide PCPP of this fact. A PCPP for a language is a proof that a given input is in the language, which can be verified by making only a few queries to both the proof and the input. In our setting, the verifier asks each prover to compute a PCPP Π for the claim that the measurement outcome string a is in the language $L = \{x : b \cdot x = c\}$. This proof can be verified by making constantly-many queries to Π, together with constantly many queries to a. Both of these correspond to *local* measurements, either of the shared quantum state, or the proof string Π generated from the measurement outcomes, and can thus be certified using our entanglement test.

There are two subtleties that arise. First, a PCPP (viewed as a nonlocal game) that is classically sound need not be sound against entangled provers. To address this, we perform a further layer of composition, encoding the PCPP proof Π in a low-degree polynomial and querying this polynomial. Secondly, in our setting *completeness* does not automatically hold either. This is because each prover j only has access to one share of the shared state, which is a qudit-by-qudit encoding of the actual QMA witness. The prover can thus only supply bits from a proof Π_j computed from its share. As a result the usual method of transforming a PCP into a game, namely by querying multiple provers for locations in the proof and checking consistency between them, fails since even honest provers do not know each other’s measurement outcomes and thus cannot answer consistently. To surmount this obstacle, we exploit the linearity of the error correcting code, together with a linear PCPP construction from [24], for which the proof Π is a linear function of the input a; the linearity holds as long as the language L is itself specified by a set of linear equations, i.e. $L = \{x : Ax = b\}$. The linearity of the PCPP allows the verifier to check consistency between one prover’s answers and the appropriate linear combination of answers returned by the other provers.[1].

With this PCPP-based protocol for measuring nonlocal Pauli observables in place, the proof of Theorem I.2 follows: starting with a QMA-hard instance of the local Hamiltonian problem with inverse-polynomial promise gap, we amplify the gap by taking a large tensor product, and then randomly sample a polynomial subset of the exponentially many terms in the tensor product. By the matrix Chernoff bound [25], with high probability this sampling preserves the promise gap, and the resulting nonlocal Hamiltonian can be tested using our protocol. (This random sampling is the source of the “randomized reductions” in Theorem I.2.)

Finally, our PCPP-based protocol enables us to check not just one nonlocal term but also many terms at once, provided that they are all tensor products of Paulis in the same basis. This allows us to obtain a protocol that accommodates an inverse-polynomial promise gap for the ground energy, provided the Hamilton is frustration free (all of its terms are simultaneously satisfied in the ground state), and each of it terms can be expressed as a tensor product of generalized τ_X or τ_Z observables, acting on an arbitrary number of qudits (see Definition 6.16 of the full version [14]). This shows point (ii) discussed earlier.

A. Techniques

Our main result, a robust entanglement test with logarithmic communication, can be stated informally as follows. For a formal statement, we refer to Theorem III.2 in Section III.

Theorem. *Let n be an integer and $q = p^t$ a prime power such that $q = \Theta(\frac{\log^2 n}{\log\log n})$. Then there exists a two-prover test* Q-LOWDEG *in which the verifier sends questions of length $\mathrm{poly}(\log n, \log q)$ and receives answers of length $O(\mathrm{poly}\log\log(n) \cdot \log q)$ such that the following hold: nolistsep*

1) (Completeness:) *There exists a strategy for the provers based on sharing an n-qudit maximally entangled state, with qudits of local dimension q, and making measurements in the eigenbasis of tensor products of generalized τ_X or τ_Z observables over $\mathbb{F}_q$;*
2) (Soundness:) *For any $\varepsilon \geq 0$, any strategy that is accepted with probability at least $1 - \varepsilon$ in the test must use an entangled state that is (up to local isometries) within distance $\delta = \mathrm{poly}(\mathrm{poly}(p) \cdot \mathrm{poly}(\varepsilon))$ from an n-qudit maximally entangled state.*[2]

A typical setting of parameters for the theorem is to choose p a constant, e.g. $p = 2$, $t = \Theta(\log\log n)$, and ε a small constant, which leads to constant soundness δ.

The test mentioned in the theorem has three components: (a) a low-degree test in the X basis; (b) a low-degree test in the Z basis; (c) an anti-commutation test relating the two bases. Both (a) and (b) are direct adaptations of the “plane-vs-point” low-degree test from [26]. The basis label, X or Z, asks the prover to measure its n qudits in the simultaneous eigenbasis of the observables $\tau_X(a)$ or $\tau_Z(b)$ defined in (1) respectively. The prover is then asked to encode the resulting outcome $a \in \mathbb{F}_q^n$ as a low-degree polynomial $g_a : \mathbb{F}_q^m \to \mathbb{F}_q$, where $m = O(\log n / \log\log n)$, and return either the

[1] We note that, just as in [24], we require linearity of the PCP in order for it to interface with a linear error correcting code.

[2] Here and throughout we use the notation $f(X) = \mathrm{poly}(h(X)))$ as an abbreviation for “there exists a universal constant $c > 0$ such that $f(X) = O(h(X)^c)$ as $X \to 0$ (if $X = \varepsilon$) or as $X \to \infty$ (if $X = n$); in the theorem p, t and q are all allowed to be implicitly functions of n, but not ε.

evaluation of the polynomial at a randomly chosen point $x \in \mathbb{F}_q^m$, or its restriction to a randomly chosen two-dimensional subspace s of $\mathbb{F}_q^m$. Part (c) is designed to enforce the "twisted commutation" relations $\tau_X(a)\tau_Z(b) = \omega^{-\mathrm{tr}(ab)}\tau_Z(b)\tau_X(a)$ satisfied by these observables. Before explaining the test and its analysis in greater detail, we first review the main steps that go into showing soundness of the classical low-degree test.

Classical low-degree tests: The effectiveness of the classical low-degree test is based on the use of the following Reed-Muller encoding of an n-variable assignment $a = (a_1, \dots, a_n) \in \{0,1\}^n$. First, integer values h and m are chosen so that $h^m \geq n$, and an injection $\pi : \{1,\dots,n\} \to \{0,\dots,h-1\}^m$ is fixed. Second, a finite field $\mathbb{F}_q$ is chosen such that $q \geq h$. Third, a function $g_a : \mathbb{F}_q^m \to \mathbb{F}_q$ is defined such that $g_a(\pi(i)) = a_i$ for all $i \in \{1,\dots,n\}$, and g_a has degree at most h in each of its m variables; g_a can be obtained by straightforward polynomial interpolation. Finally, the encoding of a is defined as the concatenation of the evaluation table of g_a at every point $x \in \mathbb{F}_q^m$ with a table describing the restriction of g_a to every two-dimensional subspace $s \subseteq \mathbb{F}_q^m$. The encoding has roughly q^{3m} entries, and each entry has size $O(d^2 \log q)$, where $d = mh$ is the total degree of g_a. Choosing $h \approx \log n$ and $m \approx \log n / \log\log n$ yields an encoding of quasi-polynomial size, $n^{O(\log n)}$, as long as q is also polynomial in n.

When used for constructions of PCPs, the low-degree test provides an encoding that can be tested and evaluated while making only a small number of queries. This is achieved based on the following observations. First, the encoding can be checked by making only a constant number of queries: the test selects a pair (x,s) such that s is a uniformly random subspace and x a uniformly random point in s, and checks consistency between the corresponding entries of the encoding. Second, the evaluation of g_a at any point $z \in \mathbb{F}_q^m$ can be recovered by making $O(d)$ queries to the encoding in a way that each query is uniformly distributed: select a uniformly random line going through z, query $d+1$ points at random on the line, and interpolate to recover the value at z.

The analysis of the low-degree test described in the previous paragraph is not simple. The goal is to show that any table which passes the test with probability $1-\varepsilon$ must be close to the encoding of a polynomial of the form g_a, for some $a \in \mathbb{F}_q^n$. The proof is constructive: it recovers a low-degree polynomial g_a through m successive steps of interpolation. The case $m = 2$ is immediate, since by definition the encoding contains the restriction of g_a to any two-dimensional subspace. For general m, one selects $(d+1)$ parallel $(m-1)$-dimensional subspaces, applies the induction hypothesis to each, and interpolates to recover a m-variate polynomial defined over the whole space. The key difficulty in the analysis is to control the error: naïvely, it would, at best, double at each step, resulting in an unmanageable blow-up. The key innovation of the test, and its analysis, is a method to limit this blow-up by a procedure of "self-improvement".

Entanglement tests: Before moving on to our quantum low-degree test, it is useful to first recall the intuition behind our prior work [17], which establishes a similar quantum analogue for the Hadamard encoding, which is based on the linearity test of Blum et al. [27].

In the linearity test, the assignment $a \in \{0,1\}^n$ is encoded as the evaluation table of the function $f_a : \mathbb{F}_2^n \to \mathbb{F}_2$, $f_a(x) = x \cdot a$. Each entry of the encoding is a single bit, but there are 2^n entries, thus the table has exponential size. The linearity test makes three queries, x, y and $x+y$ for x, y uniformly distributed in $\mathbb{F}_2^n$, and verifies that $f_a(x) + f_a(y) = f_a(x+y)$. The soundness analysis of the test is based on Fourier analysis; no induction is needed.

To turn the linearity test into a test for entanglement we first re-interpret it using the language of representation theory. The additive structure of $\mathbb{F}_2^n$ makes it into an abelian group, whose irreducible representations are the 2^n characters $\chi_a(x) = (-1)^{a \cdot x}$. An arbitrary table $f : \mathbb{F}_2^n \to \mathbb{F}_2$ can also be seen as a mapping $g = (-1)^f$ from the additive group of $\mathbb{F}_2^n$ to the 1-dimensional unitary group, $U(\mathbb{C})$. A table f which is accepted in the linearity test with probability $1-\varepsilon$ is an approximate representation of the group, in the sense that $\mathrm{E}_{x,y}\,|g(x)g(y) - g(x+y)|^2 = O(\varepsilon)$, where the expectation is uniform. Thus the analysis of the linearity test exactly amounts to showing that approximate representations of abelian groups are close to exact representations (i.e. the characters, which precisely correspond to the linear functions).

We can try to apply the same reasoning to entangled-prover strategies. Using matrix-valued Fourier analysis it is possible to show that a quantum strategy which succeeds with probability $1-\varepsilon$ in an X-basis linearity test (resp. an Z-basis linearity test) implies the existence of observables for the provers which satisfy approximate linearity conditions $X(a)X(b) \approx X(a+b)$ (resp. $Z(a)Z(b) \approx Z(a+b)$), where the approximation holds on average over uniform $a, b \in \mathbb{F}_2^n$ and is measured using the state-dependent norm that is standard in testing. These relations by themselves do not imply anything "quantum"; in particular they are satisfied by one-dimensional observables $X(a) = Z(a) = (-1)^{f(a)}$. To obtain a truly quantum test we are missing a constraint relating the two bases: the Pauli (anti)-commutation relation $X(a)Z(b) = (-1)^{a \cdot b}Z(b)X(a)$. Enforcing this relation would allow us to frame the family of unitaries $\{\pm X(a)Z(b),\, a, b \in \mathbb{F}_2^n\}$ as a representation of the Pauli group modulo complex phase (also known as the Weyl-Heisenberg group) and combine results on the stability of approximate representations [28] with information on the structure of irreducible representations of that group to conclude. This is what justifies the inclusion of part (c), an anti-commutation test, which can be based on e.g. the Mermin-Peres Magic Square game [29] to test for the desired anti-commutation relations.

A quantum low-degree test: The previous outline of an entanglement test based on the BLR linearity test is implemented in [17]. The use of the linearity test has two main advantages: (i) when executed in a single basis, its analysis with two entangled provers follows a direct argument using Fourier analysis; (ii) combining the linearity test in the X and Z bases naturally gives access to two families of observables $X(a)$ and $Z(b)$ for the provers, that can be used to specify an approximate representation of the n-qubit Weyl-Heisenberg group as described above, with the (anti)-commutation test certifying all required pairwise group relations.

To reduce the communication required in the test, it is natural to turn to low-degree tests: as described above, the latter only require poly-logarithmic, instead of linear, communication. Due to the fact that the test has only a quasi-polynomial number of questions, however, a strategy for the provers only involves a quasi-polynomial number of observables: how can one show that all exponentially many (anti)-commutation relations hold, in principle, between observables defined on the prover's space, if the test itself only requires the existence of a tiny subset of these observables in order to be played?

This difficulty can be overcome as follows. From the classical analysis of the low-degree test, or rather its entanglement-resistant analogue [12], it is possible to show that a strategy that succeeds in the X-basis (resp. Z-basis) low-degree test implies the existence of a family of observables $X(a)$ (resp. $Z(b)$), for $a \in \mathbb{F}_q^n$, that satisfy the commutation relations $X(a)X(b) = X(a+b)$ (resp. $Z(a)Z(b) = Z(a+b)$). Moreover, the use of an appropriate generalization of the Magic Square game over $\mathbb{Z}_2$, introduced in [22], to $\mathbb{Z}_s$, for any integer s, that allows us to test for the appropriate twisted commutation relation between any two observables that are actually queried in the test. The difficulty is to establish the right relations between observables $X(a)$ and $Z(b)$ that are not queried from the test, but whose existence follows from the independent application of the entangled-prover analysis of the low-degree test to the X- and Z- basis executions of the test.

Our solution proceeds in three steps. The first step consists in combining X and Z observables together into a single family of commuting observables. We do this by adjoining two ancilla systems for each prover, each initialized in a maximally entangled state local to the prover, and setting $\hat{X}(x) = X(x)_{\mathsf{A}} \otimes \tau_X(x)_{\mathsf{A'}} \otimes \mathrm{Id}_{\mathsf{A''}}$, where $\tau_X(x)_{\mathsf{A'}}$ denotes the n-qudit Pauli that the prover's $X(x)$ is supposed to implement, for x among the possible queries in the test. Defining $\hat{Z}(z)$ similarly, provided $\hat{X}(x)$ and $\hat{Z}(z)$ satisfy the (conjugate of) the twisted commutation relation satisfied by $\tau_X(x)$ and $\tau_Z(z)$ we have obtained a family of (approximately) commuting observables.

In the second step we use these commuting observables to define a strategy for the classical low-degree test, not over m-variate polynomials as the initial test requires, but over $2m$ variables, half of which are "X" variables, and half of which are "Z" variables. To construct such a strategy we have to define "points" and "subspace" measurements from the $\hat{X}(x)$ and $\hat{Z}(z)$, using the information that the initial observables $X(x)$ and $Z(z)$ came from a strategy for the provers that independently succeeded, with good probability, in the classical low-degree test. Once this has been completed we apply the analysis of the classical low-degree test against two entangled provers to recover a single family of measurements $\{\hat{S}^g\}$ with outcomes in the set of low-degree polynomials g over $\mathbb{F}_q^{2m}$.

The last step consists in "pulling apart" the measurements obtained in the previous step to recover observables $\tilde{X}(x)$ and $\tilde{Z}(z)$, now defined for all $x, z \in \mathbb{F}_q^n$ (and not only the special subset used as queries in the test). Given the definition of $\hat{X}(x)$ from $X(x)$, it is natural to define $\tilde{X} = (\mathrm{Id}_{\mathsf{A}} \otimes \mathrm{Id}_{\mathsf{A'}} \otimes \tau_X(x)_{\mathsf{A''}}) \cdot \hat{X}(x)$, which has the effect of "undoing" the initial tensoring of $X(x)$ by a Pauli on A' (this uses that the ancillas A'A" are initialized in a maximally entangled state). It remains to argue that the exponentially many operators thus constructed approximately satisfy the Pauli twisted commutation relations. Once this has been established the result follows as in our previous work [17], as it can be shown directly that such operators must be close to operators exactly satisfying all Pauli relations, whose only joint eigenvalue-1 eigenstate is the maximally entangled state.

Pauli observables over a prime power field: To conclude this overview we briefly discuss some difficulties encountered while working with generalized Pauli observables over a prime power field. Had we restricted attention to prime fields the proof (and certainly the notation!) would have been somewhat simpler. The motivation for considering prime powers comes from the desire to allow embedding qubit Hamiltonians, which we can achieve if $q = 2^t$, but did not see how to implement for odd values of q. Over prime power fields, we are faced with two possible definitions of generalized Pauli observables: the "clock" and "shift" operators mod q, with eigenvalues that are q-th roots of unity, and the definition (1), with eigenvalues that are p-th roots of unity. The former are more common in the literature and offer the convenience of allowing to encode a projective measurement with outcomes in $\mathbb{F}_q$ into a single generalized observable. However, they are not well-suited for describing strategies in the low-degree test, since they are defined in terms of addition and multiplication over $\mathbb{Z}_q$, whereas in the low degree test, all operations are performed over $\mathbb{F}_q$. Hence, we opted for the second definition, using families of t such observables to encode a single measurement with outcomes in $\mathbb{F}_q \simeq \mathbb{F}_p^t$.

B. Further work

There are several open problems raised by our work. Firstly, it would be interesting to expand the range of Hamiltonians for which we are able to give constant-gap interactive proofs, with the goal of eventually reaching a QMA-complete family, and thus a proof of Conjecture I.1 based on a deterministic reduction. Secondly, a different route towards the proof of the conjecture would consist in establishing QMA-hardness results for either of the two classes of Hamiltonians described in Definition 6.8 and Definition 6.16 of the full version [14], for which we do already have a deterministic reduction to a game. As further motivation, we note that, if such a QMA-hardness result were achieved by constructing a "history Hamiltonian" from a polynomial quantum circuit—as in all such hardness results known—then by an observation of Fitzsimons and Hajdušek [30], our results could be used to give an efficient delegation scheme for BQP in the "post-hoc" model. More broadly, the classical PCP theorem and MIP proof systems have become important tools in the design of delegated computation schemes (e.g. [31], [32]), and we hope that similar applications may arise from the games variant of QPCP. Beyond the quantum games PCP conjecture, essentially resolved in this work, the complexity of the class MIP* of languages that have multi-prover interactive proof systems with entangled provers remains wildly open. Recent work of [33] introduces a "compression" technique, that allows him to obtain MIP* protocols for language in NEEXP (non-deterministic doubly-exponential time), albeit at the cost of an exponentially small completeness-soundness gap. Could our techniques be used to obtain the same result, for a constant gap? Such a result would provide an unconditional separation between MIP and MIP*.

In a different direction, it could be useful to extend our entanglement test to sub-constant error, in the same spirit as [34], [35]. Currently, all self-testing results we are aware of only provide guarantees in a regime where the success probability is close to 1, which is arguably more challenging to demonstrate in experiments.

Organization: Due to space limitations this extended abstract is limited to an introduction of necessary notation, a description of the quantum low-degree test, and a statement of the completeness and soundness properties of the test. For intuition we also include a brief description of the honest provers' strategy in the test. The full version [14] contains the soundness analysis of the test as well as its extension to testing error-correcting code states and low-energy states of local Hamiltonians.

Acknowledgments: AN was supported by NSF CAREER Grant CCF-1452616. TV was supported by NSF CAREER Grant CCF-1553477, AFOSR YIP award number FA9550-16-1-0495, a CIFAR Azrieli Global Scholar award, and the IQIM, an NSF Physics Frontiers Center (NSF Grant PHY-1125565) with support of the Gordon and Betty Moore Foundation (GBMF-12500028). Parts of this work was completed while both authors were hosted at the Institut Henri Poincaré in Paris, as part of the special semester on Analysis in Quantum Information Theory (Fall 2017), supported by NSF Grant DMS-1700168. The hospitality of the IHP is gratefully acknowledged.

II. Preliminaries

A. Notation

We use $\mathcal{H}$ to denote a finite-dimensional Hilbert space, $\mathrm{L}(\mathcal{H})$ for the linear operators on $\mathcal{H}$, and $\mathrm{U}(\mathcal{H})$ the set of unitary operators. Subscripts $\mathcal{H}_\mathsf{A}$, $\mathcal{H}_\mathsf{B}$ indicate distinct spaces.

We use the notation $\mathrm{poly}(f(n))$ to denote $O(f^c(n))$ for some universal constant $c > 0$ (which may vary each time the notation is used). Similarly, we write $\mathrm{poly}^{-1}(f(n))$ to denote $\Omega(f^{-c}(n))$. All parameters used in the paper will generally be a function of a single parameter n, and asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, etc., should be understood as $n \to \infty$.

B. Finite fields and polynomials

Throughout we use p to denote a prime and $q = p^t$ a prime power. We let $\mathbb{F}_q$ denote the finite field with q elements, and $\mathbb{Z}_p$ denote the cyclic group mod p. The additive group of $\mathbb{F}_p$ coincides with $\mathbb{Z}_p$, but this is no longer the case for $\mathbb{F}_q$. The finite field trace is denoted by $\mathrm{tr}(a)$; it is a map from $\mathbb{F}_q$ to the prime subfield $\mathbb{F}_p$, defined by $\mathrm{tr}(a) = \sum_{\ell=0}^{t-1} a^{p^\ell}$. The trace respects linear combinations with coefficients drawn from the prime subfield: $\mathrm{tr}(\alpha a + \beta b) = \alpha \mathrm{tr}(a) + \beta \mathrm{tr}(b)$ for $\alpha, \beta \in \mathbb{F}_p$. A useful alternative view of $\mathbb{F}_q$ is as a t-dimensional vector space over $\mathbb{F}_p$. Each element $e \in \mathbb{F}_q$ can be written as $e_1 b_1 + e_2 b_2 + \cdots + e_t b_t$, where $(b_1, \ldots, b_t)$ is a basis for $\mathbb{F}_q$ over $\mathbb{F}_p$ and the coefficients e_ℓ lie in the field of scalars $\mathbb{F}_p$. This representation of $\mathbb{F}_q$ is convenient for addition, since one can add the individual components e_ℓ separately, but in general, it is hard to do multiplication. However, if q is even or $q = p^t$ with both p and t odd there always exists a basis satisfying the property of *self-duality*, i.e.

$$\mathrm{tr}(b_i b_j) = \delta_{ij} \qquad (2)$$

for all $i, j \in \{1, \ldots, t\}$ (see e.g. [36, Theorem 1.9]). This property allows to express $\mathrm{tr}(ef)$, for $e, f \in \mathbb{F}_q$, as the inner product, over $\mathbb{F}_p$, of their respective vector of components along the basis. As shown below, this property will make it convenient to express q-dimensional qudits as tensor products of p-dimensional qudits. For the remainder of the paper we only consider choices of q such that $\mathbb{F}_q$ admits a self-dual basis over $\mathbb{F}_p$.

For integer d, m and a subspace $s \subset \mathbb{F}_q^m$ we let $\deg_d(s)$ denote the set of polynomials on s of total degree at most d

(specified with respect to some fixed, implicit basis for s). We write $\omega = e^{\frac{2i\pi}{p}}$ for a fixed primitive p-th root of unity. Let

$$|\text{EPR}_q\rangle = \frac{1}{\sqrt{q}} \sum_{i \in \mathbb{F}_q} |i\rangle \otimes |i\rangle \, \in \mathbb{C}^q \otimes \mathbb{C}^q \, . \tag{3}$$

Coordinates and polynomials: Let $n \geq 1$ be an integer, and h, m two integers such that $h^m \geq n$ and $h \leq q$. Throughout we fix an arbitrary injection $\pi : \{1, \ldots, n\} \to \{0, 1, \ldots, h-1\}^m \subseteq \mathbb{F}_q^m$, where n, h, m are integers such that $h^m \geq n$ that will be clear from context. For $x \in \mathbb{F}_q^m$ and $i \in \{1, \ldots, n\}$ define

$$x_{\pi(i)} = \prod_{j=1}^{m} \frac{\prod_{\substack{k=0 \\ k \neq \pi(i)_j}}^{h-1} (k - x_j)}{\prod_{\substack{k=0 \\ k \neq \pi(i)_j}}^{h-1} (k - \pi(i)_j)} \in \mathbb{F}_q \, ,$$

and let $x_\pi = (x_{\pi(1)}, \ldots, x_{\pi(n)}) \in \mathbb{F}_q^n$. Note that for $x \in \{0, 1, \ldots, h-1\}^m$, $x_{\pi(i)} = 1$ if $x = \pi(i)$ and $x_{\pi(i)} = 0$ otherwise. By ranging over all possible values for x we obtain a subset of $\mathbb{F}_q^n$ of size q^m; we think of $x \mapsto x_\pi$ as a pseudo-random "coordinate expansion" map.

Let $g : \mathbb{F}_q^m \to \mathbb{F}_q$ be an m-variate polynomial of degree at most h in each coordinate. Then by interpolation we can write

$$g(x) = \sum_{i=1}^{n} x_{\pi(i)} g(\pi(i)) = g \cdot x_\pi \, , \tag{4}$$

where we abuse notation and write g for the vector $(g(\pi(1)), \ldots, g(\pi(n))) \in \mathbb{F}_q^n$. Conversely, for any $a \in \mathbb{F}_q^n$ we let g_a be the m-variate polynomial of individual degree at most h over $\mathbb{F}_q$ defined by

$$g_a : \, x \in \mathbb{F}_q^m \, \mapsto \, \sum_i a_i x_{\pi(i)} \, = \, a \cdot x_\pi \, . \tag{5}$$

The map from $\mathbb{F}_q^n$ to $\mathbb{F}_q^{q^m}$ that maps a to the evaluation table of g_a is the m-variate Reed-Muller code of individual degree h. Note that $(g_a(\pi(1)), \ldots, g_a(\pi(n))) = a$.

We recall the Schwartz-Zippel lemma [37], [38], which we will use repeatedly.

Lemma II.1 (Schwartz-Zippel)**.** *Let $d, m \geq 1$ be integers and r a non-zero polynomial in m variables of total degree at most d defined over the finite field $\mathbb{F}_q$. Then r has at most $d|\mathbb{F}_q|^{m-1}$ zeros.*

C. Pauli measurements and observables for qudits

To any projective measurement $\{M^a\}$ with outcomes $a \in \mathbb{Z}_p$ we can associate a generalized observable with eigenvalues that are p-th roots of unity: the unitary matrix $M = \sum_a \omega^a M^a$, where $\omega = e^{\frac{2i\pi}{p}}$. The generalized Pauli operators over $\mathbb{F}_p$ are a set of generalized observables indexed by a basis setting X or Z and an element a or b of $\mathbb{F}_p$, with eigenvalues that are p-th roots of unity. They are given by

$$\sigma_X(a) = \sum_{j \in \mathbb{F}_p} |j + a\rangle\langle j| \qquad \text{and} \qquad \sigma_Z(b) = \sum_{j \in \mathbb{F}_p} \omega^{bj} |j\rangle\langle j| \, , \tag{6}$$

where addition and multiplication are over $\mathbb{F}_p$. These observables obey the "twisted commutation" relations

$$\forall a, b \in \mathbb{F}_p, \qquad \sigma_X(a)\sigma_Z(b) \, = \, \omega^{-ab} \sigma_Z(b)\sigma_X(a) \, . \tag{7}$$

Similarly, over a field $\mathbb{F}_q$ we can define a set of generalized Pauli operators, indexed by a basis setting X or Z and an element of $\mathbb{F}_q$. There are different possible definitions for these operators. We choose them to have eigenvalues that are p-th roots of unity. For $a, b \in \mathbb{F}_q$ they are given by

$$\tau_X(a) = \sum_{j \in \mathbb{F}_q} |j + a\rangle\langle j| \quad \text{and} \quad \tau_Z(b) = \sum_{j \in \mathbb{F}_q} \omega^{\text{tr}(bj)} |j\rangle\langle j| \, ,$$

where addition and multiplication are over $\mathbb{F}_q$. Powers of these observables obey the relation

$$\forall W \in \{X, Z\}, \, \forall a \in \mathbb{F}_q, \, \forall b \in \mathbb{F}_p, \quad (\tau_W(a))^b = \tau_W(ab) \, .$$

In particular, since $pa = 0$ for any $a \in \mathbb{F}_q$ we get that that $(\tau_W(a))^p = \text{Id}$ for any $a \in \mathbb{F}_q$. The observables obey analogous "twisted commutation" relations to (7),

$$\forall a, b \in \mathbb{F}_q, \qquad \tau_X(a)\tau_Z(b) = \omega^{-\text{tr}(ab)} \tau_Z(b)\tau_X(a) \, . \tag{8}$$

It is clear from the definition that all of the τ_X operators commute with each other, and similarly all the τ_Z operators with each other. Thus, it is meaningful to speak of a common eigenbasis for all τ_X operators, and a common eigenbasis for all τ_Z operators. The common eigenbasis for the τ_Z operators is the computational basis. To map this basis to the common eigenbasis of the τ_X operators, one can apply the Fourier transform

$$F \, = \, \frac{1}{\sqrt{q}} \sum_{j,k \in \mathbb{F}_q} \omega^{-\text{tr}(jk)} |j\rangle\langle k| \, . \tag{9}$$

Explicitly, the eigenbases consist of the vectors $|e_W\rangle$ labeled by an element $e \in \mathbb{F}_q$ and $W \in \{X, Z\}$, given by

$$|e_X\rangle = \frac{1}{\sqrt{q}} \sum_j \omega^{-\text{tr}(ej)} |j\rangle \, , \qquad |e_Z\rangle = |e\rangle \, .$$

We denote the POVM whose elements are projectors onto basis vectors of the eigenbasis associated with the observables τ_W by $\{\tau_W^e\}_e$. Then the observables $\tau_W(a)$ can be written as

$$\forall W \in \{X, Z\}, \, \forall a \in \mathbb{F}_q, \qquad \tau_W(a) = \sum_{e \in \mathbb{F}_q} \omega^{\text{tr}(ae)} \tau_W^e \, .$$

For choices of q such that $\mathbb{F}_q$ admits a self-dual basis $(b_1, \ldots, b_t)$, we can decompose a q-dimensional qudit (a "quqit") as a tensor product of t p-dimensional qudits ("qupits"). Based on this decomposition, for $W \in \{X, Z\}$

and $\ell \in \{1,\dots,t\}$ we define the W-basis Pauli operator acting on the ℓ-th qupit by

$$\begin{aligned}\forall a \in \mathbb{F}_p,\ \sigma_{W,\ell}(a) \;&=\; \sum_{e_1,\dots,e_t \in \mathbb{F}_p} \omega^{a e_\ell}\, \tau_W^{(e_1 b_1 + \cdots + e_t b_t)} \\ &=\; \tau_W(a b_\ell)\;. \qquad (10)\end{aligned}$$

It can be verified by direct computation that for every $\ell \in \{1,\dots,t\}$, $\sigma_{X,\ell}$ and $\sigma_{Z,\ell}$ obey the Pauli twisted commutation relations (7), and that when $\ell \neq \ell' \in \{1,\dots,t\}$, $\sigma_{X,\ell}$ and $\sigma_{Z,\ell'}$ commute. We will sometimes consider the case where $p=2$, in which case the $\sigma_{W,\ell}$ behave as the standard Pauli spin matrices acting on t qubits, with the index ℓ labeling the qubit acted on. Also, it will be sometimes useful to allow the index a to range over all of $\mathbb{F}_q$ instead of just $\mathbb{F}_p$; extending (10) we define $\sigma_{W,\ell}(a)$ to be $\tau_W(a b_\ell)$ for any $a \in \mathbb{F}_q$.

For systems with many qudits, we will consider tensor products of the operators τ_W. Slightly abusing notation, for $W \in \{X,Z\}$ and $a \in \mathbb{F}_q^n$ we denote by $\tau_W(a)$ the tensor product $\tau_W(a_1) \otimes \ldots \otimes \tau_W(a_n)$. These obey the twisted commutation relations

$$\forall a,b \in \mathbb{F}_q^n, \qquad \tau_X(a)\tau_Z(b) \;=\; \omega^{-\mathrm{tr}(a\cdot b)}\tau_Z(b)\tau_X(a)\;,$$

where $a \cdot b = \sum_{i=1}^n a_i b_i \in \mathbb{F}_q$. For $W \in \{X,Z\}$ and $e \in \mathbb{F}_q^n$ define the eigenstates

$$|e_W\rangle = |(e_1)_W\rangle \otimes \ldots \otimes |(e_n)_W\rangle\;,$$

and associated rank-1 projectors τ_W^e.

State-dependent distance: For operators $A, B \in \mathrm{L}(\mathcal{H})$, where $\mathcal{H}$ is a finite-dimensional Hilbert space, and a vector $|\psi\rangle \in \mathcal{H} \otimes \mathcal{H}'$, where $\mathcal{H}'$ is another finite-dimensional Hilbert space, we write $A \approx_\delta B$ for $\|(A-B)\otimes \mathrm{Id}\,|\psi\rangle\|^2 = O(\delta)$. Note the state $|\psi\rangle$ and the space $\mathcal{H}'$ are usually kept implicit. We sometimes write the same with some free variables, e.g. $A_x^a \approx_\delta B_x^a$. By this we mean

$$\mathop{\mathrm{E}}_x \sum_a \|(A_x^a - B_x^a)\otimes \mathrm{Id}\,|\psi\rangle\|^2 \;=\; O(\delta)\;.$$

Variables appearing as subscript will most often be considered "inputs", and should be averaged; superscripts are considered "answers" and should be summed over. Which is which will always be clear from context, including the distribution on inputs.

D. Self-testing

We use the language of multi-player self-tests (we will often call the players "provers" as well).

Definition II.2. Let $k \geq 1$ be an integer. A *k-partite strategy* $S = (|\psi\rangle, \mathcal{X}, \mathcal{A}, \mathcal{M})$ consists of finite question and answer sets $\mathcal{X} = X_1 \times \cdots \times X_k$ and $\mathcal{A} = A_1 \times \cdots \times A_k$ respectively, a k-partite quantum state $|\psi\rangle \in \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_k$, and for each $i \in \{1,\dots,k\}$ a collection of measurement operators $\{M_x^a\}_{a \in A_i}$ on $\mathcal{H}_i$ and indexed by $x \in X_i$.[3]

We reproduce a standard definition in self-testing.

Definition II.3. A *k-player self-test* with completeness c and robustness $\delta(\varepsilon)$ for a strategy $S = (|\Psi\rangle, \mathcal{X}, \mathcal{A}, \mathcal{M})$ is a distribution π on $\mathcal{X}$ and a family of coefficients $V(a_1,\dots,a_k|x_1,\dots,x_k) \in [0,1]$, for $(x_1,\dots,x_k) \in \mathcal{X}$ and $(a_1,\dots,a_k) \in \mathcal{A}$, such that the following hold: nolistsep

- Strategy S and succeeds in the test with probability at least c;
- Any strategy $\hat{S} = (|\hat{\psi}\rangle, \mathcal{X}, \mathcal{A}, \hat{\mathcal{M}})$ with success at least $c - \varepsilon$ in the test must be $\delta(\varepsilon)$-close to the optimal strategy. Formally, there exists a local isometry $\Phi = \Phi_1 \otimes \cdots \Phi_k$ and a state $|\mathrm{AUX}\rangle$ such that

$$\big\|\Phi(|\hat{\psi}\rangle) - |\mathrm{AUX}\rangle|\psi\rangle\big\| \;\leq\; \delta(\varepsilon)\;,$$

and

$$\begin{aligned}\mathop{\mathrm{E}}_{x_1,\dots,x_k \sim \pi} \sum_{a_1,\dots,a_k} &\big\|\Phi\big(\hat{M}_{x_1}^{a_1} \otimes \cdots \otimes \hat{M}_{x_k}^{a_k}|\hat{\psi}\rangle\big) \\ &- |\mathrm{AUX}\rangle M_{x_1}^{a_1} \otimes \cdots \otimes M_{x_k}^{a_k}|\psi\rangle\big\| \leq \delta(\varepsilon)\;.\end{aligned}$$

In case S only specifies a partial strategy, then the above expression is restricted to questions for which S is defined.

E. The commutation test

In designing self-tests, it is useful to have the ability to test commutation relations between pairs of observables applied by the provers. The following well-known test can be employed to certify that two observables commute:

Theorem II.4. *Let s be an integer and $\varepsilon > 0$. There exists a two-player self-test* $\mathrm{COM}(M,N)$ *with completeness* 1 *and robustness $\delta(\varepsilon) = O(s\sqrt{\varepsilon})$, for the strategy S that uses commuting generalized observables M and N (with outcomes in $\mathbb{Z}_s$) for two special questions labelled* 1 *and* 2*, respectively. The test has* 3 *questions per player and answers either in $\mathbb{Z}_s$ (for questions* 1 *and* 2*) or $\mathbb{Z}_s^2$ (for question* 3*). Moreover, for any two commuting observables A and B, there exists a strategy in which the first player uses the observables M and N for questions* 1 *and* 2*, using a shared state $|\psi\rangle$ that is a maximally entangled state of appropriate dimension.*

The guarantees of the theorem are achieved by the following test, which is a simple instance of the idea of "oracularization" in multiprover interactive proofs. In the test, the verifier performs either of the following with equal probability $\frac{1}{2}$:

1) Send the first player a question q chosen uniformly from $\{1,2\}$, and send the second player the question 3. Receive an answer $a \in \mathbb{Z}_s$ from the first player

[3] Although this is left implicit in the notation, the measurement operators associated with different spaces need not be equal.

and $(b_1, b_2) \in \mathbb{Z}_s^2$ from the second player. Accept if $a = b_q$, and reject otherwise.

2) Perform the same as in item 1., but with the players interchanged.

For the analysis of this test see e.g. [39, Lemma 28].

F. The generalized Magic Square

In [22] a generalized version of the Magic Square game [29] is introduced and shown to robustly self-test generalized observables satisfying twisted commutation relations over $\mathbb{Z}_s$, for any integer s.

Theorem II.5 (Theorem 5.9 in [22])**.** *Let s be an integer and $\varepsilon > 0$. There exists a two-player self-test* $\mathrm{MS}(X, Z)$*, with completeness* 1 *and robustness $\delta(\varepsilon) = O(s^3\sqrt{\varepsilon})$, for the (partial) strategy S that uses observables σ_X and σ_Z on two special questions labeled X and Z respectively. The test has $O(1)$ questions per player (including two questions labeled X and Z) and answers in $\mathbb{Z}_s^2$. Furthermore, there is a strategy that succeeds with probability* 1 *using only σ_X, σ_Y and σ_Z observables on two s-dimensional qudits per player initialized in $|\psi\rangle = |\mathrm{EPR}_s\rangle \otimes |\mathrm{EPR}_s\rangle$.*

G. The classical low-degree test

A stepping stone in our analysis is an extension of the "classical low-degree test" from [10] to the case of only two provers.

Theorem II.6 (Theorem 2 in [12])**.** *Let $\varepsilon > 0$, m, d integers, and q a prime power such that $q \geq (dm/\varepsilon)^c$ for a universal constant $c \geq 1$. There is a two-prover test, called the* classical low-degree test C-LOWDEG(m, d, q)*, in which queries to the provers are chosen among affine subspaces $s \subseteq \mathbb{F}_q^m$, and answers are polynomials r on s of total degree at most d, such that the following holds. For any strategy for the provers using entangled state $|\psi\rangle$ and projective measurements $\{M_s^r\}$ that succeeds with probability at least $1 - \varepsilon$ in the test there exists a POVM $\{S^g\}$, where g ranges over the polynomials on $\mathbb{F}_q^m$ of total degree at most d, and a $\delta = \mathrm{poly}(\varepsilon)$ such that the following hold:*

1) *Approximate consistency with M:*

$$\mathop{\mathrm{E}}_{s} \sum_{g} \sum_{r \neq g|_s} \langle\psi| M_s^r \otimes S^g |\psi\rangle \leq \delta,$$

2) *Self-consistency:*

$$\sum_g \langle\psi| S^g \otimes (\mathrm{Id} - S^g) |\psi\rangle \leq \delta.$$

We let π_{ld} denote the distribution on questions used by the verifier in the low-degree test from Theorem II.6. This distribution is symmetric, and we slightly abuse notation by also writing π_{ld} for either marginal. We will use that the test from Theorem II.6 that it satisfies the following properties: nolistsep

(i) π_{ld} is a uniform mixture of the uniform distribution on pairs (s, w) such that s is an affine subspace of dimension 2 in $\mathbb{F}_q^m$ and $w \in s$ is a uniformly random point in s, and its permutation (w, s).

(ii) Whenever provers in the test are queried for a pair of subspaces (s, w), they are required to return a polynomial r defined on s and a value a in $\mathbb{F}_q$ such that $r(w) = a$.

The length of questions in the low-degree test C-LOWDEG(m, d, q) from Theorem II.6 is $O(m \log q)$, which for a choice of $q = \mathrm{poly}\log(n)$ is logarithmic in n. However, answers have length $O(d^2 \log q)$, which is super-logarithmic. To achieve reduced answer length it is standard to compose the test with itself, and we denote the resulting test by C-LOWDEG$^{(2)}(m, d, q)$. In the composed test, any answer r from a prover is interpreted as an $n' = O(d^2 \log q)$-long string of bits, that can be encoded as a multilinear polynomial over $\mathbb{F}_q^{m'}$, for m' such that $2^{m'} \geq n'$. Questions in the composed test are a subspace $s \subseteq \mathbb{F}_q^m$, together with a subspace $s' \subseteq \mathbb{F}_q^{m'}$, and answers are the restriction to s' of the low-degree encoding of the polynomial r that the prover would answer to the question s. We refer the reader to the full version for a precise statement of soundness for the composed test.

III. The quantum low-degree test

We denote our quantum low-degree test by Q-LOWDEG$^{(l)}$, for $l \in \{1, 2\}$. Here l denotes the "level" of the test, before ($l = 1$) or after ($l = 2$) composition. In general we also write Q-LOWDEG for the "composed quantum low-degree test" Q-LOWDEG$^{(2)}$, which is the variant of the test with reduced answer size, and is the variant that will be used in our applications. The test is described in Figure 1. We show that the test is a self-test for the following class of Pauli strategies. To define the strategy, recall the definition of the POVM $\{\tau_W^a\}$ in Section II-C, defined for each $W \in \{X, Z\}$. For $s \subset \mathbb{F}_q^m$ either a point or a 2-dimensional subspace, and r a polynomial defined on s, define

$$\tau_{W,s}^r = \sum_{a \in \mathbb{F}_q^n :\, (g_a)_{|s} = r} \tau_W^a\,, \tag{11}$$

where g_a is defined in (5). Finally, for reasons that will become clear later, it is convenient to introduce

$$\boldsymbol{\tau}_{X,s}^r = \tau_{X,s}^{-r} \qquad \text{and} \qquad \boldsymbol{\tau}_{Z,s}^r = \tau_{Z,s}^r\,. \tag{12}$$

Definition III.1. Let p be a prime, $t \geq 1$ an integer, and $q = p^t$. The low-degree Pauli strategy S_{P} on n qudits of local dimension q is the strategy $(|\psi\rangle, \mathcal{X}, \mathcal{A}, \mathcal{M})$ where $|\psi\rangle = |\mathrm{EPR}_q\rangle^{\otimes n}$, $\mathcal{X} = \{X, Z\} \times (\mathcal{X}_1 \cup \mathcal{X}_2)$, where $\mathcal{X}_1 = \mathbb{F}_q^m$ and $\mathcal{X}_2$ is the set of all two-dimensional subspaces of $\mathbb{F}_q^m$, $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$, where $\mathcal{A}_1 = \mathbb{F}_q$ and $\mathcal{A}_2 = \deg_d(\mathbb{F}_q^2)$, and $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$, where $\mathcal{M}_1 = \{\tau_{W,w}^a\} \times \{\boldsymbol{\tau}_{W,w}^a\}$ and

Test Q-LOWDEG$^{(l)}(m,d,q)$. m,d are integer, and $q = p^t$ is a prime power such that $\mathbb{F}_q$ admits a self-dual basis $(b_1,\dots,b_t)$ over $\mathbb{F}_p$.The verifier performs the following with equal probability:

(a) Select $W \in \{X,Z\}$ uniformly at random and send W to both provers. If $l = 2$ execute the composed low-degree test C-LOWDEG$^{(2)}(m,d,q)$ with the provers. If $l = 1$ execute the test C-LOWDEG(m,d,q) from Theorem II.6. Let r be the polynomial returned by the first prover, and r' by the second. If $W = X$, set $A = r$ and $\boldsymbol{A'} = -r'$. If $W = Z$, set $A = r$ and $\boldsymbol{A'} = r'$. Accept if and only if the pair of answers $(A, \boldsymbol{A'})$ would have been accepted in the classical test.

(b) Select $x,z \in \mathbb{F}_q^m$ and $u,u' \in \mathbb{F}_q$ uniformly at random, and let $a = \mathrm{tr}((ux_\pi)\cdot(u'z_\pi)) \in \mathbb{F}_p$.

- If $a = 0$, execute the self-test COM (see Theorem II.4), replacing queries 1, 2, and 3 in the test by (X,x), (Z,z), and (x,z,uu') respectively, and in the case of queries 1 and 2, replacing the prover's answer $b \in \mathbb{F}_q$ by $\mathrm{tr}(ub)$ or $\mathrm{tr}(u'b) \in \mathbb{F}_p$, respectively, before making the same decision as the verifier in the test.
- If $a \neq 0$, execute the self-test MS (see Theorem II.5) with the following modification: the question labeled X is replaced by the query (X,x) as in part (a), and the prover's answer $b \in \mathbb{F}_q$ is replaced by $\mathrm{tr}(ub) \in \mathbb{F}_p$; the question labeled Z is replaced by the query (Z,z) as in part (a), and the prover's answer $b \in \mathbb{F}_q$ is replaced by $a^{-1}\mathrm{tr}(u'b) \in \mathbb{F}_p$.

Figure 1. The quantum low-degree test. $l \in \{1,2\}$ denotes the "level" of the test, before ($l = 1$) or after ($l = 2$) composition.

$\mathcal{M}_2 = \{\tau^r_{W,s}\} \times \{\boldsymbol{\tau}^r_{W,s}\}$, with $\tau^a_{W,w}$, $\tau^r_{W,s}$, and $\boldsymbol{\tau}^a_{W,w}$, $\boldsymbol{\tau}^r_{W,s}$ defined as in (11) and (12) respectively.

Theorem III.2. *Let $n \geq 1$ be an integer. Let h,m be integer such that $h^m \geq n$, and let $d = hm$. Let $q = p^t$ be a prime power such that $\mathbb{F}_q$ admits a self-dual basis over $\mathbb{F}_p$. Then for any $\varepsilon \geq 0$ the test* Q-LOWDEG$^{(2)}(m,d,q)$ *is a 2-prover self-test for the low-degree Pauli strategy S_P on n qudits of local dimension q with completeness 1 and robustness $\delta =$* $\mathrm{poly}(\mathrm{poly}(p)\cdot\mathrm{poly}(\varepsilon) + \mathrm{poly}(d/q))$. *Moreover, the test has questions of length $O(m\log q)$ and answers of length $O(\log^2(d)\log(q))$.*

Remark III.3. In a typical application of the test Q-LOWDEG$^{(2)}$, the parameters are chosen such that $m = \Theta(\frac{\log n}{\log\log n})$ and $h = \Theta(\log n)$, resulting in $d = \Theta(\frac{\log^2(n)}{\log\log n})$. Further, we chose p to be constant and $q = \Theta(\frac{\log^2(n)}{\log\log n})$ such that d/q is a small constant. This results in a question length that is $O(\log n)$ and an answer length that is $\mathrm{poly}(\log\log n)$.

The proof of the following lemma specifies the "honest" strategy of the provers in the quantum low-degree test.

Lemma III.4 (Completeness)**.** *For m,d,q as in Theorem III.2 the strategy S_P introduced in Definition III.1 can be extended to a strategy that succeeds with probability* 1 *in the test* Q-LOWDEG(m,d,q).

Proof: Let $|\psi_{\mathrm{EPR}}\rangle = \otimes_{j=1}^{n+1}|\mathrm{EPR}_q\rangle$, where $|\mathrm{EPR}_q\rangle$ is defined in (3). We first describe a strategy for the players assuming questions in part (a) of the test come from C-LOWDEG, instead of the composed test C-LOWDEG$^{(2)}$. Once a strategy for the former has been defined it is straightforward to adapt it to a strategy for the latter; this only requires classical post-processing.

To define the strategy we use the generalized Pauli operators and projections defined in Section II-C. When queried for a subspace $s \subseteq \mathbb{F}_p^m$ in a basis $W \in \{X,Z\}$, the prover measures the first n qudits using the projective measurement $\{\tau_W^a\}$ and returns the polynomial $(g_a)_{|s}$; this corresponds to the POVM described in (11).

It remains to show that these measurements define a strategy which succeeds with probability 1 in both parts of the test. We refer to the full version [14] for details. ■

References

[1] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, "Proof verification and the hardness of approximation problems," *J. ACM*, vol. 45, no. 3, pp. 501–555, 1998.

[2] S. Arora and S. Safra, "Probabilistic checking of proofs: A new characterization of NP," *J. ACM*, vol. 45, no. 1, pp. 70–122, 1998.

[3] D. Aharonov, I. Arad, and T. Vidick, "The quantum PCP conjecture," Tech. Rep., 2013, appeared as guest column in ACM SIGACT News archive Volume 44 Issue 2, June 2013, Pages 47–79.

[4] D. Aharonov, I. Arad, Z. Landau, and U. Vazirani, "The detectability lemma and quantum gap amplification," in *Proc. 41st STOC*. New York, NY, USA: ACM, 2009, pp. 417–426.

[5] D. Aharonov and L. Eldar, "The commuting local Hamiltonian problem on locally expanding graphs is approximable in NP," *Quantum Information Processing*, vol. 14, no. 1, pp. 83–101, Jan. 2015.

[6] F. G. Brandao and A. W. Harrow, "Product-state approximations to quantum ground states," in *Proc. 45th STOC*, 2013.

[7] L. Eldar and A. W. Harrow, "Local Hamiltonians whose ground states are hard to approximate," 2015.

[8] J. Fitzsimons and T. Vidick, "A multiprover interactive proof system for the local Hamiltonian problem," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 2015, pp. 103–112.

[9] L. Babai, L. Fortnow, and C. Lund, "Non-deterministic exponential time has two-prover interactive protocols," *Computational Complexity*, vol. 1, pp. 3–40, 1991.

[10] T. Vidick, "Three-player entangled XOR games are NP-hard to approximate," in *Proc. 54th FOCS*, 2013.

[11] T. Ito and T. Vidick, "A multi-prover interactive proof for NEXP sound against entangled provers," *Proc. 53rd FOCS*, pp. 243–252, 2012.

[12] A. Natarajan and T. Vidick, "Two-player entangled games are NP-hard," 2017, to appear in the proceedings of CCC'18.

[13] Z. Ji, "Classical verification of quantum proofs," in *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2016, pp. 885–898.

[14] A. Natarajan and T. Vidick, "Low-degree testing for quantum states, and a quantum entangled games PCP for QMA," 2018. [Online]. Available: https://arxiv.org/abs/1801.03821v2

[15] M. Bavarian, T. Vidick, and H. Yuen, "Hardness amplification for entangled games via anchoring," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2017, pp. 303–316.

[16] A. B. Grilo, I. Kerenidis, and A. Pereszlényi, "Pointer quantum PCPs and multi-prover games," 2016.

[17] A. Natarajan and T. Vidick, "A quantum linearity test for robustly verifying entanglement," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2017. New York, NY, USA: ACM, 2017, pp. 1003–1015. [Online]. Available: http://doi.acm.org/10.1145/3055399.3055468

[18] R. Chao, B. W. Reichardt, C. Sutherland, and T. Vidick, "Test for a large amount of entanglement, using few measurements," in *Proceedings of the 2017 Conference on Innovations in Theoretical Computer Science (ITCS)*, 2017.

[19] D. Ostrev and T. Vidick, "Entanglement of approximate quantum strategies in XOR games," Tech. Rep., 2016.

[20] M. Coudron and A. Natarajan, "The parallel-repeated Magic Square game is rigid," Tech. Rep., 2016.

[21] A. W. Coladangelo, "Parallel self-testing of (tilted) EPR pairs via copies of (tilted) CHSH," Tech. Rep., 2016.

[22] A. Coladangelo and J. Stark, "Robust self-testing for linear constraint system games," 2017.

[23] T. Cubitt and A. Montanaro, "Complexity classification of local Hamiltonian problems," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 120–129.

[24] E. Ben-Sasson, O. Goldreich, P. Harsha, M. Sudan, and S. Vadhan, "Short pcps verifiable in polylogarithmic time," in *Computational Complexity, 2005. Proceedings. Twentieth Annual IEEE Conference on*. IEEE, 2005, pp. 120–134.

[25] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 569–579, 2002.

[26] R. Raz and S. Safra, "A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP," in *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, ser. STOC '97. New York, NY, USA: ACM, 1997, pp. 475–484. [Online]. Available: http://doi.acm.org/10.1145/258533.258641

[27] M. Blum, M. Luby, and R. Rubinfeld, "Self-testing/correcting with applications to numerical problems," *Journal of Computer and System Sciences*, vol. 47, pp. 549–595, 1993.

[28] W. T. Gowers and O. Hatami, "Inverse and stability theorems for approximate representations of finite groups," 2015.

[29] P. K. Aravind, "The magic squares and Bell's theorem," Tech. Rep., 2002.

[30] J. Fitzsimons and M. Hajdušek, "Post hoc verification of quantum computing," Tech. Rep., 2015.

[31] Y. T. Kalai, R. Raz, and R. D. Rothblum, "How to delegate computations: The power of no-signaling proofs," in *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, ser. STOC '14. New York, NY, USA: ACM, 2014, pp. 485–494. [Online]. Available: http://doi.acm.org/10.1145/2591796.2591809

[32] O. Reingold, G. N. Rothblum, and R. D. Rothblum, "Constant-round interactive proofs for delegating computation," in *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, ser. STOC '16. New York, NY, USA: ACM, 2016, pp. 49–62. [Online]. Available: http://doi.acm.org/10.1145/2897518.2897652

[33] Z. Ji, "Compression of quantum multi-prover interactive proofs," 2016.

[34] R. Arnon-Friedman and H. Yuen, "Noise-tolerant testing of high entanglement of formation," 2017.

[35] R. Arnon-Friedman and J.-D. Bancal, "Device-independent certification of one-shot distillable entanglement," 2017.

[36] A. J. Menezes, I. F. Blake, X. Gao, R. C. Mullin, S. A. Vanstone, and T. Yaghoobian, *Applications of finite fields*. Springer Science & Business Media, 2013, vol. 199.

[37] R. Zippel, "Probabilistic algorithms for sparse polynomials," in *Proceedings of the International Symposiumon on Symbolic and Algebraic Computation*, ser. EUROSAM '79. London, UK, UK: Springer-Verlag, 1979, pp. 216–226. [Online]. Available: http://dl.acm.org/citation.cfm?id=646670.698972

[38] J. T. Schwartz, "Fast probabilistic algorithms for verification of polynomial identities," *Journal of the ACM*, vol. 27, no. 4, pp. 701–717, 1980.

[39] A. Coladangelo, A. Grilo, S. Jeffery, and T. Vidick, "Verifier-on-a-leash: new schemes for verifiable delegated quantum computation, with quasilinear resources," 2017.