Quantum Information and Black Holes

Jared Barron*and David Wandler†

December 7, 2018

1 Introduction

One of the most important problems in physics is finding a theory that can completely describe both quantum mechanics and gravity. The most prominent physical example of a system where quantum fields meet strong gravitational fields is near black holes. One of the first and most important studies of quantum field theory around black holes is Hawking's paper [1] showing that, under some reasonable assumptions about the nature of quantum gravity, black holes should give off radiation, leading to their eventual evaporation. This result prompted an intense discussion on how to deal with quantum information in the vicinity of a black hole that persists to this day. The purpose of this report is to highlight some interesting connections between the fields of black hole physics and quantum information theory.

We start by discussing the seemingly simple problem of whether it is possible to recover quantum information that falls into a black hole. To facilitate this discussion, we describe some physics results about what happens to the infalling radiation and how it interacts with the Hawking radiation that the black hole emits. This is done in Section 2. Section 3 then describes how Hayden and Preskill [2] transformed the way physicists think about this problem by casting it into the terms of quantum computing as a task of unscrambling quantum information, and presents their argument that this task is information-theoretically possible. We then present in Section 4 a more recent result from Kitaev and Yoshida [3] that describes an algorithm to solve this computational problem.

Our discussion then pivots to an example of how quantum information theory can lead to surprising new physics and give us valuable insight into the nature of quantum gravity. We begin this discussion in Section 5 by summarizing the AMPS firewall paradox [4]. In Section 6 we describe how Hayden and Harlow [5] were able to avoid this paradox using quantum information theory. Section 7 ends with a brief discussion of the current situation in the firewall paradox debate.

2 Physics

In modern physics, gravity is described by general relativity which specifies the geometry of spacetime. In particular, it defines how all kinds of matter and radiation are able to move between points in spacetime. For example, it tells us that nothing can travel faster than the speed of light. In this language, a black hole can be informally defined as a spatially bounded region from which no object or signal can escape. Perhaps surprisingly, general relativity permits the existence of such regions, and we have observed them.

The fact that nothing can escape from a black hole makes them seem very useful for getting rid of unwanted information. Let us say that Alice has some information that she wants to make sure that Bob can never get his hands on. All she has to do is throw it into a black hole, and there will be no way for Bob to extract it out of the black hole and into the rest of the universe.

The problem with this argument is that it is completely classical. What would happen if we introduce quantum mechanics to this scenario? Would Bob be able to access this information with the right equipment?

*jared.barron@mail.utoronto.ca
†f.wandler@mail.utoronto.ca

To answer this question rigorously, we would need a complete theory of quantum gravity, but some interesting progress has been made with some fairly reasonable assumptions on what quantum gravity might be like.

Firstly, we need to know how black holes behave when we bring quantum mechanics to the table. In 1974, Hawking published a paper [1] describing that a black hole interacting with quantum fields should give off radiation. To see why this happens, we need to consider how we count particles in quantum field theory. In quantum field theory, the objects that we are concerned with are fields, which are operator-valued functions of spacetime. The operators that we define using quantum fields are like the operators in quantum circuits in that they act on quantum states, though they need not be unitary. It is an important fact that these operators can be decomposed into operators that raise and lower the energy of a particular state by a particular finite amount. These operators are known as creation and annihilation operators, respectively, and the discrete, finite amounts of energy that they create and destroy are the particles of particle physics. Therefore, in order to count particles you must have access to the creation and annihilation operators of the field theory.

Modern gravitational theory allows for changes in the geometry of spacetime. The details of how this change in geometry affects and is affected by quantum fields is the central problem for finding a theory of quantum gravity; however, it is reasonable to assume that for ordinary circumstances (such as far away from the singularity at the center of a black hole) quantum field theory stays the same, just with a different underlying spacetime. This is the semiclassical approximation of quantum gravity and is justified by the fact that here on Earth, away from strong gravitational fields, quantum field theory appears to work. The difficulty with this description is that the change of spacetime changes the fields (which were functions of spacetime) and makes it more difficult to decompose them into creation and annihilation operators. To overcome this problem, Hawking looked at two areas of the black hole geometry that were essentially flat: the distant past before the black hole forms and the distant future far away from the black hole. In these two regions of spacetime, there is no issue finding creation and annihilation operators and using them to count particles. The brilliant part of the argument is that the creation and annihilation operators in the two regions are different, meaning the number of particles that are counted for a particular quantum state can be different in the two regions. If we start with a zero particle vacuum before the black hole forms, then we could still see particles later on. This is particle generation by the black hole. In particular, Hawking did some detailed calculations to show that the number and energies of particles created by the black hole is consistent with an object radiating because of heat. This lead to the Hawking temperature of the black hole:

$$T_H = \frac{\hbar c^3}{8\pi G k_B M} \tag{1}$$

Here M is the mass of the black hole, G is Newton's gravitational constant, k_B is Boltzmann's constant, $\hbar$ is Planck's constant, and c is the speed of light in a vacuum. For our purposes, the exact form of the temperature will not be important, but the result is too interesting to resist writing down!

Relating this back to our original discussion of Alice disposing of information into a black hole, the obvious question arises: does the radiation contain the information of what went in? In other words, if Bob were extremely clever could he manipulate this radiation in such a way as to tell what information Alice threw in?

At first, the answer seems to be no. If something has fallen into the black hole it can no longer interact with the surface or the radiation being emitted from it. However, this introduces fundamental problems from a quantum mechanical point of view. Suppose we have 3 qubits in the state $\frac{1}{\sqrt{2}}|000\rangle + \frac{1}{\sqrt{2}}|011\rangle$ and we let the first two qubits fall into the black hole. This is fine and does not change the state, we just can no longer access the first two qubits. However, the black hole will be radiating and radiation carries energy; this means that the energy of the black hole must decrease. For a simple black hole (i.e. one that is not rotating) the only energy the black hole has to give is its mass, so the mass of the black hole must decrease. Eventually, it will get to the point where there is no mass left and the black hole will vanish completely. What happens to our quantum state at this point? The first two qubits have simply vanished along with the black hole and their information cannot have come out with the radiation, so we are left with the mixed state $\rho = \frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|1\rangle\langle 1|$. This means that the natural process of the black hole absorbing information and evaporating has changed a pure state into a mixed state. It is a mathematical fact that this cannot

be done by an unitary operator, which contradicts the rule in quantum mechanics that natural processes must be unitaries (i.e. unitarity). Recall that unitarity is why we can only use gates that perform unitary transformations.

This contradiction tells us that either the idea of unitarity must be wrong or the idea that quantum information cannot come out of a black hole must be wrong. Susskind, Thorlacius, and Uglum in 1993 [6] provided a way to keep unitarity and have the Hawking radiation contain the information of what fell into the black hole. Their argument is very technical and uses a considerable amount of general relativity, but the main conclusions are fairly accessible. Essentially, they find that describing the state in the interior of the black hole and the state outside the black hole cannot be done by one Hilbert space. Instead, anyone falling into the black will see states described by a certain Hilbert space that is different than the Hilbert space of observers that stay outside the black hole. This allows them to define a membrane near the edge of the black hole that can only be described in the Hilbert space of the outside observers. This membrane contains all the information of what fell into the black hole and mixes it up according to complicated unitary transformations, then emits it in the Hawking radiation. Any observer falling into the black hole will not encounter this membrane, however, since it cannot be described in their Hilbert space, but this disagreement can never be expressed since it is impossible for them to communicate their findings to the observers outside the black hole. In this way, matter can fall into the black hole without anything strange happening, but outside observers still see the information come out and behave according to unitary operations.

This gives some hope for Bob, though Susskind et al do point out that the information will be so horribly scrambled that it would only be accessible by looking at correlations in the radiation emitted over long time scales.

3 The Problem of Decoding Hawking Radiation

Armed with the results of the previous section, physicists started looking into how the information comes out of the black hole. Hayden and Preskill took a nice approach by representing the information in the problem as qubits, meaning that the black hole stores some number of qubits and the radiation comes out as qubits. They found that if a black hole has already radiated away over half of its qubits, then a k-qubit state thrown into the black hole will likely be recoverable once only a few more than k qubits have been emitted [2]. To show this we need to recall some facts about entanglement and mixed states.

Suppose that you have 2 subsystems of qubits, A with m qubits and B with n qubits. If $m \leq n$ then you can have A maximally entangled with some subsystem of B. This means that the state of the system is given by

$$\frac{1}{\sqrt{2^m}} \sum_{x \in \{0,1\}^m} |x\rangle_A \otimes |f(x)\rangle_B$$

where $f : \{0,1\}^m \to \{0,1\}^n$ is some injective function. In this case, the mixed state representation for the system A is

$$\frac{1}{\sqrt{2^m}} I$$

where I is the $2^m \times 2^m$ identity matrix. This is known as the maximally mixed state. On the other hand, if there is no entanglement between A and B then we can write the whole system as

$$|\psi\rangle_A \otimes |\phi\rangle_B .$$

In fact, even if there are more subsystems, the condition for no entanglement between A and B is

$$\rho_{AB} = \rho_A \otimes \rho_B$$

where ρ_A, ρ_B, and ρ_{AB} are density matrices for subsystem A, subsystem B, and the combined subsystem AB, respectively. Also relevant is the notion of entropy of a quantum subsystem. For subsystem A with density matrix ρ_A, the von Neumann entropy is defined as

$$S_A = -\operatorname{tr}(\rho_A \ln \rho_A) .$$

Notice that for the maximally mixed state of the m-qubit subsystem, the entropy can be easily evaluated as $\ln(2^m)$. This is the maximum entropy that a quantum state can have and can only be reached by a maximally mixed state. Using these facts, we are ready to follow Hayden and Preskill's argument.

Suppose that Alice throws a k-qubit quantum state into a black hole that contains (on its "information membrane") $n-k$-qubits. We also suppose that the qubits in the black hole are maximally entangled with the m-qubits of Hawking radiation that have already left the black hole. This assumption may seem a little dubious, but Don Page [7] was able to show that if $1 << n-k < m$ and we take an average over all pure states of the radiation and the black hole combined, the average entropy in the black hole state is

$$\ln\left(2^{n-k}\right) - \frac{n-k}{2m}$$

This means that as we increase m the average black hole state gets very close to the maximally mixed state. The maximal mixed state could only arise in this case if the black hole is maximally entangled with the already emitted radiation. This justifies the claim that our assumption is reasonable.

To follow Alice's secret information, we suppose that she had her k-qubit state maximally entangled with another k-qubit state that remains outside the black hole. Then it becomes sufficient to show that Bob can generate a state that is maximally entangled with this k-qubit reference system. This is essentially just a bit of book keeping; it makes it so that we can talk about the information in Alice's state relying only on the entanglement instead of keeping track of every individual qubit.

Now with all of that setup, we finally let Alice throw her k-qubits into the black hole. Before the scrambling, we have $n-k$-qubits in the black hole that are entangled with the previously emitted radiation and k-qubits that are entangled with the reference system. For simplicity, we will follow Hayden and Preskill in labeling the early radiation E, the reference system N, and the combined n-qubit black hole system B. The scrambling then acts on B with some n-qubit unitary operator V. We have no idea what V will actually look like apart from an n-qubit unitary matrix; to deal with this, we will eventually average over all possible matrices and say that the result applies to a typical operator V. In any case, after the scrambling, we are left with an n-qubit black hole that is maximally entangled with the combined subsystem NE. After some time, the scrambled B will have split into a new $n-s$-qubit black hole B' and s qubits of radiation which we label R.

Now we suppose that Bob is free to use anything that comes out of the black hole, so he has access to subsystems E and R. Then, all we need to show is that N is now maximally entangled with the combined system RE. This is equivalent to saying that it is no longer entangled with the black hole, B'. Therefore, it is sufficient to show that for the average V, $\rho_{NB'}(V)$ is close to $\rho_N(V) \otimes \rho_{B'}(V)$. These states are given by

$$\rho_{NB'}(V) = \mathrm{tr}_R\left((I_N \otimes V)\rho_{NB}(I_N \otimes V^\dagger)\right)$$

and

$$\rho_N(V) = \mathrm{tr}_{RB'}\left((I_N \otimes V)\rho_{NB}(I_N \otimes V^\dagger)\right)$$

where I_N is the identity operator on the subsystem N and tr_A denotes the *partial trace with respect to A*. The partial trace with respect to A just means find the mixed state without the subsystem A. Since the left over black hole B' must be maximally entangled with the radiation by our previous assumptions, we take it to be the maximally mixed state, $\rho_{B'}(V) = 2^{s-n} I_{B'}$. To show our desired result, we use the L_1 norm[1] to quantify closeness and integrate over all the possible V using the Haar measure normalized to one. Hayden and Preskill use a mathematical identity to bound this quantity[2]:

$$\int dV ||\rho_{NB'}(V) - \rho_N(V) \otimes 2^{s-n} I_{B'} \leq \frac{2^{n+k}}{2^{2s}} \mathrm{tr}\left(\rho_{NB}^2\right). \tag{2}$$

Now, recall that we had B was made up of $n-k$ qubits that were maximally entangled to E and k qubits that were maximally entangled to N. Therefore, $\rho_{NB} = 2^{k-n} I_{2^{n-k}} \otimes |\psi\rangle\langle\psi|$ where $|\psi\rangle$ describes the $2k$-qubit

[1] The L_1 norm for matrices is given by $||A|| = \mathrm{tr}\sqrt{A^\dagger A}$

[2] The justification of this identity is beyond the scope of this report. The curious reader should look at the references of Hayden and Preskill [2].

fully entangled state with Alice's qubits and the subsystem N. Therefore,

$$\mathrm{tr}\left(\rho_{NB}^2\right) = \mathrm{tr}\Big(2^{2(k-n)} I_{2^{n-k}} \otimes |\psi\rangle\langle\psi|\Big) = 2^{k-n}$$

Hence, we have

$$\int dV ||\rho_{NB'}(V) - \rho_N(V) \otimes 2^{s-n} I_{B'} \leq 2^{2(k-s)}. \tag{3}$$

Notice that if Bob waits for only a few more than k qubits to come out of the black hole, the average scrambling operator V will scramble the information in such a way that the entanglement with N is all transferred to the radiation. Therefore, if Bob could collect the radiation into a powerful enough quantum computer he could decode it to get a k-qubit state with the same entanglement properties as Alice's original message. In the next section, we discuss an algorithm to do this decoding, assuming Bob knows enough quantum gravity to know V.

4 The Solution to Decoding Hawking Radiation

Here we offer a quantum computing algorithm that is able to decode the Hawking radiation as described above, given a few simplifying assumptions. This algorithm is based entirely on that proposed by Kitaev and Yoshida [3], though we have simplified the setup for easier discussion. The analysis of the algorithm is based on EPR states. Now we know that the EPR state for 2 qubits is $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. This can be generalized to the $2n$ qubit EPR state:

$$\left|EPR^{2n}\right\rangle = \frac{1}{\sqrt{2^n}} \sum_{x\in\{0,1\}^n} |x\rangle |x\rangle \tag{4}$$

These states have very useful and unique properties. For example, for any n-qubit operators A and B we have:

$$(A \otimes B)\left|EPR^{2n}\right\rangle = (I \otimes A^T B)\left|EPR^{2n}\right\rangle. \tag{5}$$

In particular, if U is an n-qubit unitary operator, then

$$(U \otimes U^*)\left|EPR^{2n}\right\rangle = (I \otimes U^T U^*)\left|EPR^{2n}\right\rangle = (I \otimes I)\left|EPR^{2n}\right\rangle = \left|EPR^{2n}\right\rangle. \tag{6}$$

This property also holds if we replace $\left|EPR^{2n}\right\rangle$ with $\left|EPR^{2(n-k)}\right\rangle\left|EPR^{2k}\right\rangle$, that is

$$(U \otimes U^*) \sum_{x\in\{0,1\}^{n-k}} \sum_{y\in\{0,1\}^k} |x\rangle |y\rangle |x\rangle |y\rangle = \sum_{x\in\{0,1\}^{n-k}} \sum_{y\in\{0,1\}^k} |x\rangle |y\rangle |x\rangle |y\rangle \tag{7}$$

To prove this one merely needs to split up U into block matrices and make repeated use of Equation 5.

For the algorithm to work as described by Kitaev and Yoshida, we use a few assumptions. The first and perhaps least reasonable assumption is that Bob has processed the information in E to produce an $n-k$-qubit system $\tilde{B}$, such that B and $\tilde{B}$ form the state $\left|EPR^{2(n-k)}\right\rangle$. Based on our discussion in the previous section, this is possible, but it has been argued that the computation required to generate this state takes an exponentially long time [5][3]. Next we assume that the state Alice throws in is half of a $2k$-qubit EPR state and that Bob knows the size of Alice's input, has access to an idealized quantum computer, and is able to implement V^* and V^T. This last assumption on Bob should be reasonable if Bob knows enough quantum gravity theory to predict V. They also assume that the radiation leaves the black hole in the maximally mixed state[4].

Based on the above assumptions, the initial state of the system before Alice throws her qubits into the black hole is

$$|\Psi_{init}\rangle = \frac{1}{\sqrt{2^k 2^{n-k}}} \sum_{y\in\{0,1\}^k} \sum_{x\in\{0,1\}^{n-k}} |y\rangle |x\rangle |y\rangle |x\rangle \tag{8}$$

[3] We discuss this paper more in Section 6.

[4] In reality we expect the thermally mixed state, but this is more difficult to deal with than the maximally mixed state and is not covered in [3]

This is just the tensor product of the 2 EPR states written in a convenient way. The first k-qubits correspond to the reference system N, the next $n-k$ qubits to the black hole B, the next k to Alice's qubits, and the last $n-k$ to Bob's processed radiation $\tilde{B}$. Then Alice throws her qubits into the black hole and the scrambling unitary V is applied to everything in the black hole. This leads to the state

$$(I \otimes V \otimes I) |\Psi_{init}\rangle = \frac{1}{\sqrt{2^k 2^{n-k}}} \sum_{y \in \{0,1\}^k} \sum_{x \in \{0,1\}^{n-k}} |y\rangle V (|x\rangle |y\rangle) |x\rangle . \tag{9}$$

To solve the decoding problem, Bob must find some unitaries that he can apply to only the last $n-k+s$ qubits of this state, plus any ancilla qubits he needs, to reliably find k qubits in the EPR state with N (i.e. the first k qubits). For simplicity in our description we take $s = k$. Bob starts by introducing $2k$ ancilla qubits in the EPR state. He then applies the unitary V^* to the n qubit system made up of $\tilde{B}$ and the first half of the ancilla state. This gives an overall state of

$$\begin{aligned} &\frac{1}{\sqrt{2^{n+k}}} \sum_{z \in \{0,1\}^k} \sum_{y \in \{0,1\}^k} \sum_{x \in \{0,1\}^{n-k}} |y\rangle V (|x\rangle |y\rangle) V^* (|x\rangle |z\rangle) |z\rangle \\ &\quad = (I \otimes V \otimes V^* \otimes I) \frac{1}{\sqrt{2^{n+k}}} \sum_{z \in \{0,1\}^k} \sum_{y \in \{0,1\}^k} \sum_{x \in \{0,1\}^{n-k}} |y\rangle |x\rangle |y\rangle |x\rangle |z\rangle |z\rangle \end{aligned} \tag{10}$$

Consider now the state without $I \otimes V \otimes V^* \otimes I$ applied. This state is just the tensor product of 3 EPR states. For any $2k$-subsystem, we can rewrite this state with respect to an orthonormal basis containing the state $|ERP_{2k}\rangle$. Doing this to the subsystem made up of Alice's qubits and the first half of the ancillary qubits we see that after some rearranging:

$$\begin{aligned} &\frac{1}{\sqrt{2^{n+k}}} \sum_{z \in \{0,1\}^k} \sum_{y \in \{0,1\}^k} \sum_{x \in \{0,1\}^{n-k}} |y\rangle |x\rangle |y\rangle |x\rangle |z\rangle |z\rangle \\ &\quad = \frac{1}{\sqrt{2^{n+k}}} \left(\frac{1}{\sqrt{2^k}} \sum_{z \in \{0,1\}^k} \sum_{y \in \{0,1\}^k} \sum_{x \in \{0,1\}^{n-k}} |y\rangle |x\rangle |z\rangle |x\rangle |z\rangle |y\rangle + |\Phi\rangle \right) \end{aligned} \tag{11}$$

Here $|\Phi\rangle$ is some state of the system which is orthogonal to any state which has the described subsystem in the $2k$-qubit EPR state. Notice two important properties of the first term in this expansion: the reference system is in the EPR state with the last k ancillary qubits and, by Equation 7, it is invariant under the operator $I \otimes V \otimes V^* \otimes I$. Hence, the final state after applying V^* is:

$$\frac{1}{2^k \sqrt{2^n}} \sum_{z \in \{0,1\}^k} \sum_{y \in \{0,1\}^k} \sum_{x \in \{0,1\}^{n-k}} |y\rangle |x\rangle |z\rangle |x\rangle |z\rangle |y\rangle + (I \otimes V \otimes V^* \otimes I) \frac{1}{\sqrt{2^{n+k}}} |\Phi\rangle . \tag{12}$$

The state $(I \otimes V \otimes V^* \otimes I) \frac{1}{\sqrt{2^{n+k}}} |\Phi\rangle$ might produce terms that with the radiation and first ancilla bits in the EPR state which would hinder the use of this algorithm. However, if we use our assumption that the radiation leaves the black hole in a mixed state, then there must be an equal contribution of all basis vectors in any orthogonal basis for the radiation and first ancilla subsystem. Any significant term of $(I \otimes V \otimes V^* \otimes I) \frac{1}{\sqrt{2^{n+k}}} |\Phi\rangle$ with this subsystem in the EPR state will add to the first term of the equation and skew this state to being too likely to be consistent with the maximally mixed assumption. Therefore, we can say that the only term with the radiation and first k ancilla bits in the EPR state is the first term. Hence, if Bob measures the radiation and the first k ancillary qubits in the orthonormal basis described above and obtains the EPR state, then the state collapses to

$$\frac{1}{\sqrt{2^{n+k}}} \sum_{z \in \{0,1\}^k} \sum_{y \in \{0,1\}^k} \sum_{x \in \{0,1\}^{n-k}} |y\rangle |x\rangle |z\rangle |x\rangle |z\rangle |y\rangle \tag{13}$$

and Bob has successfully completed his decoding. This is what Kitaev and Yoshida call the "probabilistic decoder", but the probability that it actually returns the desired decoded state is 2^{-k}.

Fortunately, Kitaev and Yoshida found a clever trick to overcome this obstacle. Notice that if we call the state in Equation 13 $|\psi\rangle$ and define $\sqrt{1-2^{-k}}\,|\phi\rangle \equiv (I \otimes V \otimes V^* \otimes I)\frac{1}{\sqrt{2^{n+k}}}\,|\Phi\rangle$. The overall state can then be written as

$$\sqrt{2^{-k}}\,|\psi\rangle + \sqrt{1-2^{-k}}\,|\phi\rangle = \sin\frac{\theta}{2}\,|\psi\rangle + \cos\frac{\theta}{2}\,|\phi\rangle \tag{14}$$

Here we have used $\sin\left(\frac{\theta}{2}\right) \equiv \sqrt{2^{-k}}$, to illustrate the similarity between this setup and the setup for Grover's Algorithm. In fact, Kitaev and Yoshida introduce an iterate that performs the same kind of rotation as the Grover iterate. To do this they use the projection operators

$$\begin{aligned} P_A =& (I_{n+k} \otimes V^* \otimes I_k)(I_{2n} \otimes |EPR_{2k}\rangle\langle EPR_{2k}|)(I_{n+k} \otimes V^T \otimes I_k) \\ P_D =& \frac{1}{\sqrt{2^k}} I_n \otimes |x\rangle\langle x| \otimes I_{n-k} \otimes |x\rangle\langle x| \otimes I_k \end{aligned} \tag{15}$$

Notice that these operators leave the first n qubits of our state unchanged, which means Bob is free to implement them. Then the iterate is given by $W = (2P_A - I)(I - 2P_D)$ and each application of the iterate increases the argument of the sine and cosine in Equation 14 by exactly θ. Therefore, after applying W the correct number of times ($\mathcal{O}(2^k)$) Bob will measure the desired state with high probability. Thus, this algorithm gives a satisfactory solution to the problem of decoding Alice's information.

5 The Firewall Paradox

Black hole complementarity had seemingly resolved the information paradox, under conditions that most agreed were reasonable; first, that a black hole evolves unitarily according to an S-matrix description of quantum mechanics. Second, that beyond a microscopic distance away from the black hole horizon, semi-classical quantum field theory could approximate physics well. Third, that the dimension of the space of states of a black hole is given by the exponential of its entropy. And fourth, that an observer falling into a black hole experiences no 'drama'; that is, they encounter no unduly high-energy quanta. However, in 2013 a paper authored by Almheiri, Marolf, Polchinski, and Sully (AMPS) upset the status quo once more by arguing that these postulates are inconsistent [4]. They posited that an observer falling into an old black hole would either encounter incredibly energetic particles (a firewall) at the black hole horizon, or that non-local effects, not describable by current quantum field theory, would extend to a macroscopic distance beyond the event horizon. If the black hole is old, the subspace of early radiation has much larger dimension than that of the late radiation, so it is highly probable for a full state of Hawking radiation to be in the form

$$\Psi = \sum_i |\psi_i\rangle_E \otimes |i\rangle_L$$

where $|i\rangle_L$ is an orthogonal basis for the late radiation subspace, and $\langle\psi_i|\psi_j\rangle \propto \delta_{ij}$. That is, it is highly probable for the late radiation to be almost maximally entangled with the early radiation. The essence of the argument is that because the Hawking radiation is postulated to be in a pure state, and the late radiation is almost maximally entangled with the early radiation, an observer measuring the state of early Hawking radiation could predict the number of particles at a given energy in the late radiation. Explicitly, the observer can measure in the number basis for the late, outgoing radiation with the operator $b^\dagger b$, which is valid everywhere outside the black hole. When Hawking radiation is observed at low energy far away from the black hole, one can relate it to a blue-shifted (higher energy) mode near the black hole - radiation is red-shifted by the distortion in spacetime after it is emitted from the black hole. The assumption that the infalling observer experiences no drama is equivalent to saying that if they measure in the number-operator basis for infalling Hawking modes, with lowering operator a, they will measure the vacuum state. AMPS showed that this is inconsistent with the observer's hypothetical ability to measure in the number basis of the b operator. So, the observer will not see a vacuum of infalling modes, they will see a high-energy 'firewall'. AMPS also make a case for their paradox using entropy of entanglement, which will be more familiar for those with a background in quantum information. Quantum state entropy follows a law of strong sub-additivity [8]. For a system with sub-systems A, B, and C:

$$S_{AB} + S_{BC} \geq S_B + S_{ABC}$$

Recall that S_A gives a measure of entanglement between A and the other part of the system, BC. If we take A as the qubits emitted early in the black hole's lifetime, B as a late qubit emitted, and C as the 'partner' qubit falling into the black hole, we have the following conditions. Because the black hole is old and losing mass, equivalent to losing entropy, we can say that $S_{AB} < S_A$. Because the infalling observer experiences no drama, which is conditional on the late outgoing qubit and its infalling partner being maximally entangled, $S_{BC} = 0$. Now, the following two inequalities also hold for general subsystems A and B [9]:

$$S_{AB} \leq S_A + S_B$$

$$S_{AB} \geq |S_A - S_B|$$

Applying this to S_A and S_{BC}, as subsystems of S_{ABC}, yields that $S_{ABC} = S_A$. Substituting the various entropies into our statement of subadditivity, we have

$$S_A \geq S_{AB} = S_{AB} + S_{BC} \geq S_B + S_A$$

However, since the late radiation by itself is thermal, its partial density matrix is mixed, and has positive entropy. So, this inequality must be violated, which is a contradiction. The essence of the paradox is that the observer can measure entanglement between early and late Hawking radiation, then jump in the black hole and confirm entanglement between the late radiation and its infalling partner. But this violates the monogamy of entanglement, and effectively the no-cloning theorem. On the other hand, if the entanglement between the late radiation and its infalling partner is broken, the observer sees a firewall at the black hole's horizon, and the 'no-drama' postulate is violated. The argument for firewalls is highly nuanced and reliant on technical details of quantum field theory. Its introduction has led to intense scrutiny as to whether the circumstances leading to the paradox can be avoided. Within a year, Hayden and Harlow published a paper making the case that firewalls can be avoided if the confirmation of entanglement between early and late radiation is too computationally complex, and arguing that such a computation is indeed very likely to be computationally difficult. We shall follow their argument, and explore the consequences for current research on the relationship between quantum information and black holes.

6 Quantum Computation Defeats the Firewall Paradox

For a Schwarzschild black hole of mass M, composed of n qubits, the entropy is proportional to M^2, and evaporation time is proportional to M^3. So, a freefalling observer must process information from $n \sim M^2$ bits in $\mathcal{O}(M^3) \sim n^{3/2}$ time. Hayden and Harlow presented a convincing argument that the required computation is likely exponential in n [5]. Some might argue that performing the computation used by AMPS is irrelevant to the physical reality of the firewall, but since the introduction of complementarity, the principle of operationalism, that only those measurements that can actually be made are relevant to determine physical reality, has reigned in the area of black hole physics. The AMPS experiment is recast as an explicit quantum computing problem in the following way:
Consider a finite basis for the radiation field outside the black hole given by states

$$|b_1 b_2 \dots b_k, h_1 \dots h_m, r_1 \dots r_{n-k-m}\rangle_R$$

There are k bits near the horizon and m still in the black hole. Note that these are not bits in the black hole, but are labeled in correspondence with them. Label the bits in the black hole the subsystem H, the late radiation still near the black hole B, and R the early radiation. Then, for an old black hole we can write the full state as

$$|\Psi\rangle = \frac{1}{\sqrt{|B||H|}} \sum_{b,h} |b\rangle_B |h\rangle_H U_R |b, h, 0\rangle_R$$

That is, the radiation bits are entangled with the bits in the hole and near the horizon, but this entanglement is obscured by some scrambling unitary operator U_R whose nature depends on the details of quantum gravity. The AMPS computational task is to effectively find U_R^{-1} so that the entanglement can be confirmed. Now, imagine that Alice has some set of qubits that comprise her quantum computer C, that she prepares in a

particular initial state. After the quantum computation, we wish to have the k bits still in the black hole entangled fully with the first k bits of the quantum computer's memory. Hayden and Harlow fix the time the computer runs for, so the combined state of the system and quantum computer evolve under some fixed unitary operator U_{comp}, which is determined by the local physics. The initial state $|\Psi\rangle_C$ of the quantum computer should satisfy:

$$U_{comp}(U_R |b, h, 0\rangle_R \otimes |\Psi\rangle_C) = |\text{anything}\rangle_{R,C\setminus 1,2,\ldots k} \otimes |b\rangle_{1,2,\ldots k}$$

Of course this computation will be approximate, so we cast an ϵ-net over the Hilbert space, which is of cardinality $\mathcal{O}((\frac{2}{\epsilon})^{2|R||C|})$. There are $(\frac{2}{\epsilon})^{2|R||C|2^{-k}}$ possible states for the $|\text{something}\rangle$, and 2^{k+m} possible states for $|b, h, 0\rangle$, so for a random U_{comp}, the probability that a given $|\Psi_C\rangle$ yields a desired outcome for all possible initial states of radiation is

$$\left(\left(\frac{2}{\epsilon}\right)^{2|R||C|2^{-k}}\left(\frac{2}{\epsilon}\right)^{-2|R||C|}\right)^{-2^{k+m}} = \left(\frac{2}{\epsilon}\right)^{2|R||C|2^m(2^k-1)}$$

With $(\frac{2}{\epsilon})^{2|C|}$ possible initial states, the probability that Alice chooses a correct one is $(\frac{2}{\epsilon})^{2|C|(|R|2^m(2^k-1)-1)}$ This scales inversely with a double exponential in k. Hayden and Harlow comment that even by searching over varying size $|C|$ of quantum computer, this probability does not improve. By varying the runtime to sample different unitary operators, she will expect to find an answer in $\sim 10^{10^{10^{40}}}$ years; effectively waiting to land on a correct result by complete chance, letting the quantum system evolve through all possible states. The time to perform the computation scales as $\sim e^{2\log(\frac{2}{\epsilon})|R||C||B||H|}$. By using a quantum gate circuit model, the double exponential can be reduced to a single exponential, through an application of the Solovay-Kitaev theorem. Because U_R sits in an equivalence class with freedom in its action on the $n-k-m$ last bits in R, Hayden and Harlow finally conclude that the time to complete the computation is $\sim 2^{k+n+m}$. This clearly out scales any small polynomial in n, so Alice will not have time to complete the AMPS experiment before the black hole evaporates, eliminating the possibility of encountering a firewall. Further subtleties regarding any clever choice Alice could make, or preparation of the black hole, were not found to significantly change this result.

This problem can also be cast into the language of quantum error-correcting codes. The entanglement of the radiation with the black hole acts as an erasure channel, and the scrambling U_R as an encoding. In general, even for small encoding circuits, error correction requires exponentially sized error-correcting circuits. We will now introduce the theorem that HH used to justify their assertion that Alice's computation is very difficult. First, we must introduce the quantum complexity class called Quantum Statistical Zero Knowledge ($QSZK$). In this class of problems, there is a verifier with polynomial time resources and access to a secretive, omnipotent prover. The prover can always succeed on problems for which the answer is 'yes' but will fail with some probability otherwise, and after several queries can convince the verifier of the answer, without the verifier ever knowing anything about how the problem is solved. Since the verifier has polynomial resources, $BQP \subseteq QSZK$ trivially. The following problem is complete in $QSZK$: For three systems B, H, and R, and some polynomial-sized circuit U such that $|\psi\rangle_{BHR} = U|000\rangle_{BHR}$ has maximal entanglement between B and HR, determine whether the entanglement can be decoded only from R. Now, if Alice can efficiently perform the decoding whenever it is possible, she can certainly determine whether the decoding can be done, so the above problem would be in BQP, implying $BQP = QSZK$. In other words, an omnipotent prover with access to a quantum computer would be no help at all over using a quantum computer for polynomial time. Hayden and Harlow argue that this would be extremely surprising, so we should believe it is not true. This argument, while not a proof, is in the spirit of much of complexity theory, where the evidence for many conjectures about hierarchies of complexity classes relies on the inability of humans so far to find counterexamples.

7 Or does it?

Just as the introduction of black hole complementarity did not prevent the argument for firewalls, and just as firewalls did not prevent a computational argument against them, Hayden and Harlow's decoding task

did not end the debate on black hole information. Since 2013, new ideas both for and against firewalls have been proposed[10][11]. In 2014, Unruh and Oppenheim made the case that one could create a black hole, already entangled with a quantum memory, and even before creating the black hole, perform the required computation on the quantum memory [10]. This protocol goes as follows: Create an initial state

$$|\Psi\rangle = \frac{1}{\sqrt{|B||H|}} \sum_{b,h} |b\rangle_{M_B} |h\rangle_{M_H} W_{BH} |bh\rangle_{BH}$$

where M_B and M_H are registers of the quantum memory, $|bh\rangle_{BH}$ is the matter that will form the black hole, and W_{BH} is the unitary process that describes the formation and evolution of the black hole. This is clearly a maximally entangled state. Now, after a long time, some particles are emitted from the black hole, which is modeled to be a unitary process U_{BH}. We call this subsystem of emitted particles B, and since the black hole partial is maximally entangled with the quantum memory, its partial density matrix is essentially maximally mixed, so we can say that $\|\rho_{BH} - \rho_H \otimes \rho_B\| \leq \epsilon$. That is, H and B are very close to being in a product state. Now, since M_B and M_H were the purifications of B and H, there exists some unitary V_M acting only on the quantum memory such that

$$V_M \otimes U_{BH} |\Psi\rangle = \frac{1}{\sqrt{|B|}} \sum_{b} |b\rangle_{M_{\bar{B}}} |b\rangle_B \otimes \frac{1}{\sqrt{|H|}} \sum_{h} |h\rangle_{M_{\bar{H}}} |h\rangle_H$$

for some subsystems $M_{\bar{H}}$ and $M_{\bar{B}}$ of M_{HB}. The entanglement of this state can be easily verified. Since the unitary operation V_M that decodes the entanglement does not need to be applied to the Hawking radiation emitted by the black hole, it can be found and applied before the black hole forms, removing the time constraint used by Hayden and Harlow. In effect, the Hawking radiation comes out already decoded. Alice is then free to jump into the black hole and verify entanglement with the infalling particles, showing violation of monogamy of entanglement, or otherwise encountering a firewall. Of course, this argument too has its weaknesses and possible points of failure. The debate on firewalls is far from settled, and sits squarely at the intersection between research in quantum gravity and quantum information theory.

8 Conclusion

The study of black hole information has been ongoing for almost 50 years, and in that time our understanding of the problem has undergone many radical revolutions, from Hawking's information loss, to Susskind's complementarity, to firewalls and beyond. Hopefully this brief tour through the history of the subject has highlighted the important concepts, and impressed upon the reader the remarkably deep connection between purely gravitational physics and quantum information theory. Absent a general theory of quantum gravity, quantum computing methods show great promise in being able to effectively model the processes involved in black hole evolution, the scrambling of information, and the decoding of entanglement required to make sense of the Hawking radiation emitted by black holes.

References

[1] S. W. Hawking. Particle creation by black holes. *Communications in Mathematical Physics*, 43(3):199–220, Aug 1975.

[2] Patrick Hayden and John Preskill. Black holes as mirrors: quantum information in random subsystems. *Journal of High Energy Physics*, 2007(09):120, 2007.

[3] Beni Yoshida and Alexei Kitaev. Efficient decoding for the Hayden-Preskill protocol. 2017.

[4] Ahmed Almheiri, Donald Marolf, Joseph Polchinski, and James Sully. Black holes: complementarity or firewalls? *Journal of High Energy Physics*, 2013(2):62, Feb 2013.

[5] Daniel Harlow and Patrick Hayden. Quantum computation vs. firewalls. *Journal of High Energy Physics*, 2013(6):85, Jun 2013.

[6] Leonard Susskind, Lárus Thorlacius, and John Uglum. The stretched horizon and black hole complementarity. *Phys. Rev. D*, 48:3743–3761, Oct 1993.

[7] Don N. Page. Average entropy of a subsystem. *Phys. Rev. Lett.*, 71:1291–1294, Aug 1993.

[8] Elliott H. Lieb and Mary Beth Ruskai. Proof of the strong subadditivity of quantum-mechanical entropy. *Journal of Mathematical Physics*, 14(12):1938–1941, 1973.

[9] Huzihiro Araki and Elliott H. Lieb. Entropy inequalities. *Communications in Mathematical Physics*, 18(2):160–170, Jun 1970.

[10] Jonathan Oppenheim and Bill Unruh. Firewalls and flat mirrors: An alternative to the amps experiment which evades the harlow-hayden obstacle. *Journal of High Energy Physics*, 2014(3):120, Mar 2014.

[11] Ning Bao, Adam Bouland, Aidan Chatwin-Davies, Jason Pollack, and Henry Yuen. Rescuing complementarity with little drama. *Journal of High Energy Physics*, 2016(12):26, Dec 2016.