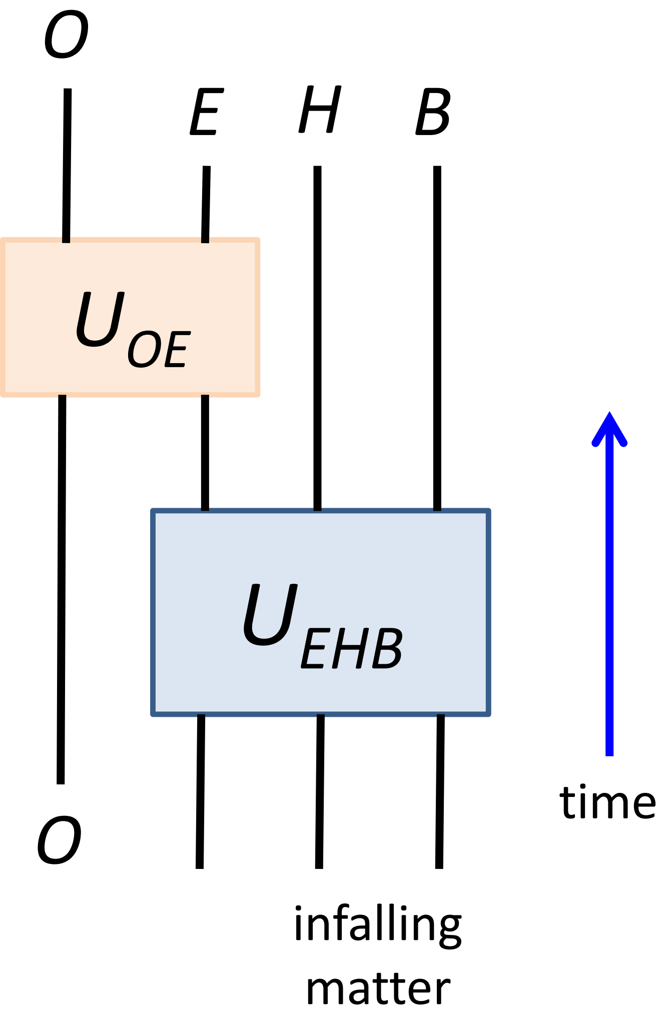
The Ghost in the Radiation: Robust encodings of the black hole interior

Isaac Kim

Eugene Tang



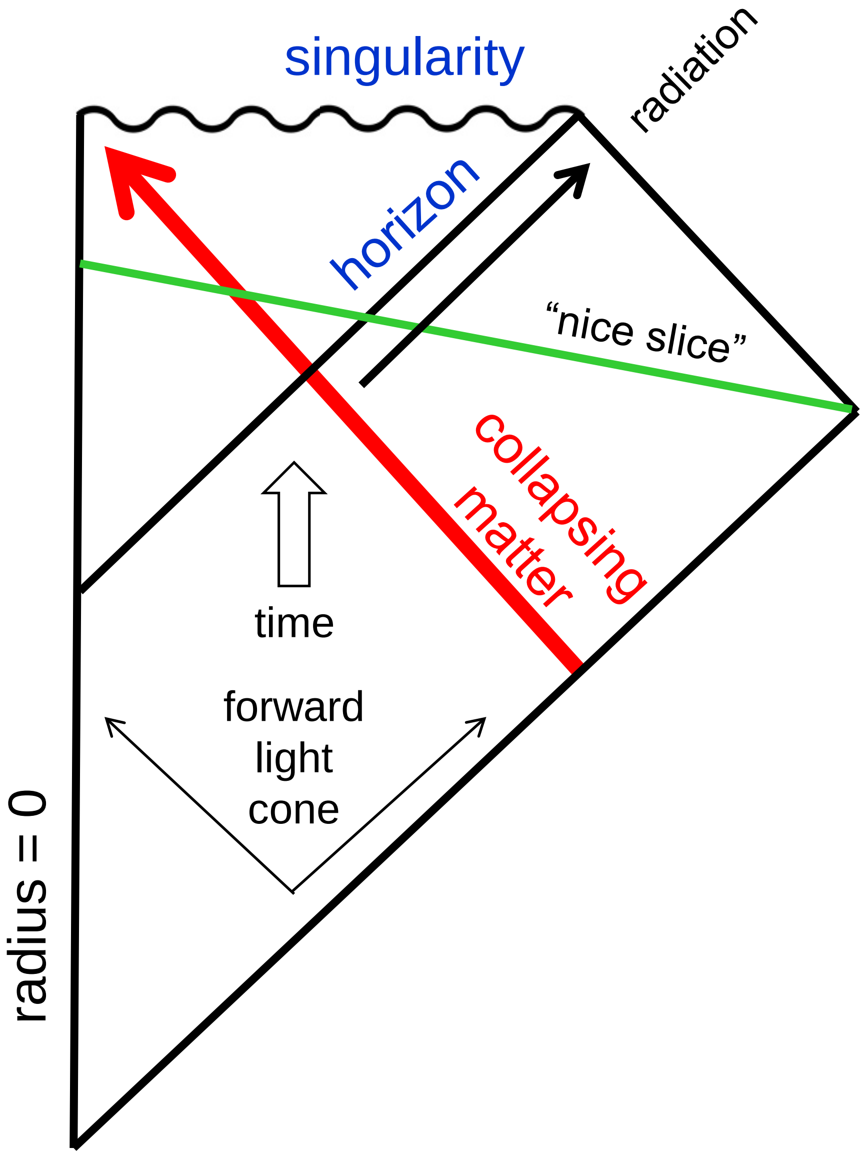
INSTITUTE FOR QUANTUM INFORMATION AND MATTER

John Preskill

Geometry from the Quantum

KITP, 17 January 2020

Black hole as quantum cloner



If information escapes from an evaporating black hole, then …

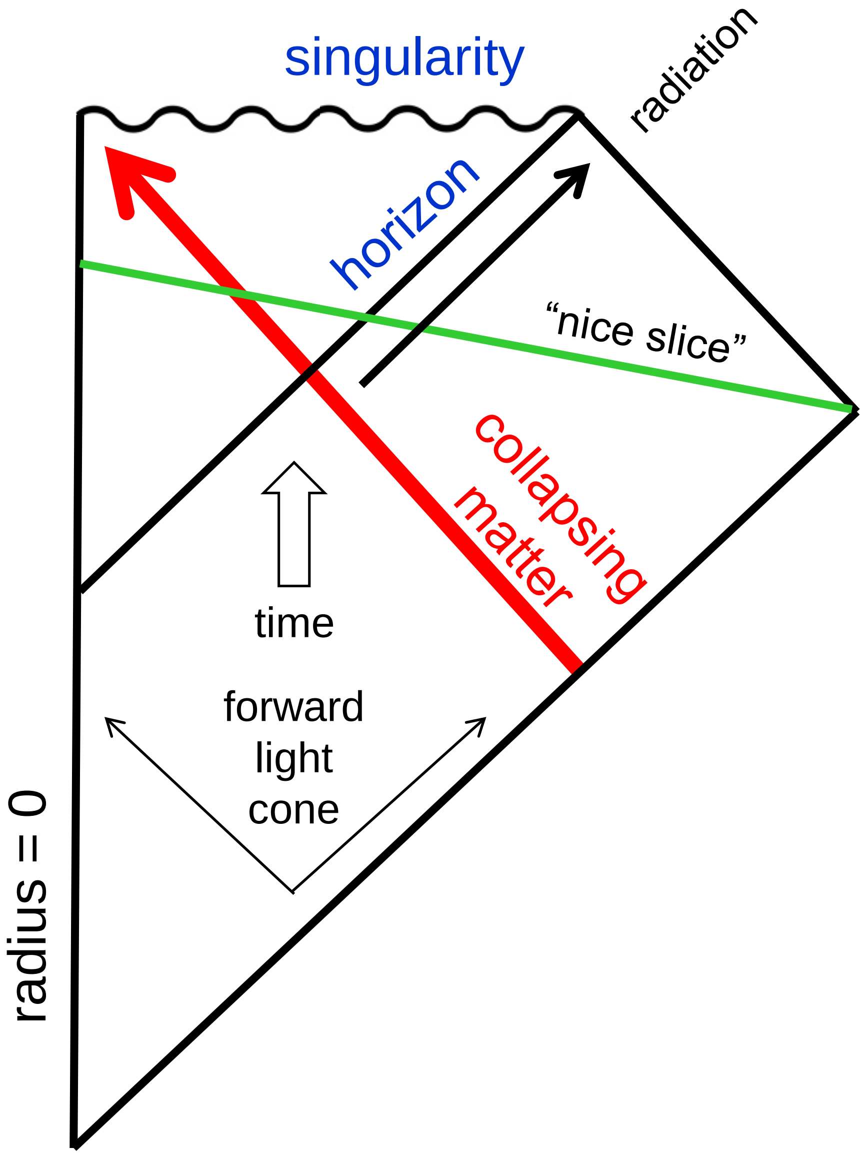
The same quantum information is in two different places at the same time.

The "nice slice" shown in green can be chosen to cross both the collapsing body behind the horizon and 99% of the escaping Hawking radiation outside the horizon.

Yet the slice only occupies regions of low curvature, where we would normally expect semiclassical physics to be reliable.

We're stuck: Either information is destroyed or cloning occurs. Either way, quantum physics needs revision.

“Black hole complementarity”



The inside and the outside are not two separate systems.

$$\mathcal{H} \neq \mathcal{H}_{\mathrm{in}} \otimes \mathcal{H}_{\mathrm{out}}$$

Rather, they are two different ways of looking at the *same* system. [Susskind 1993].

Complementarity Challenged

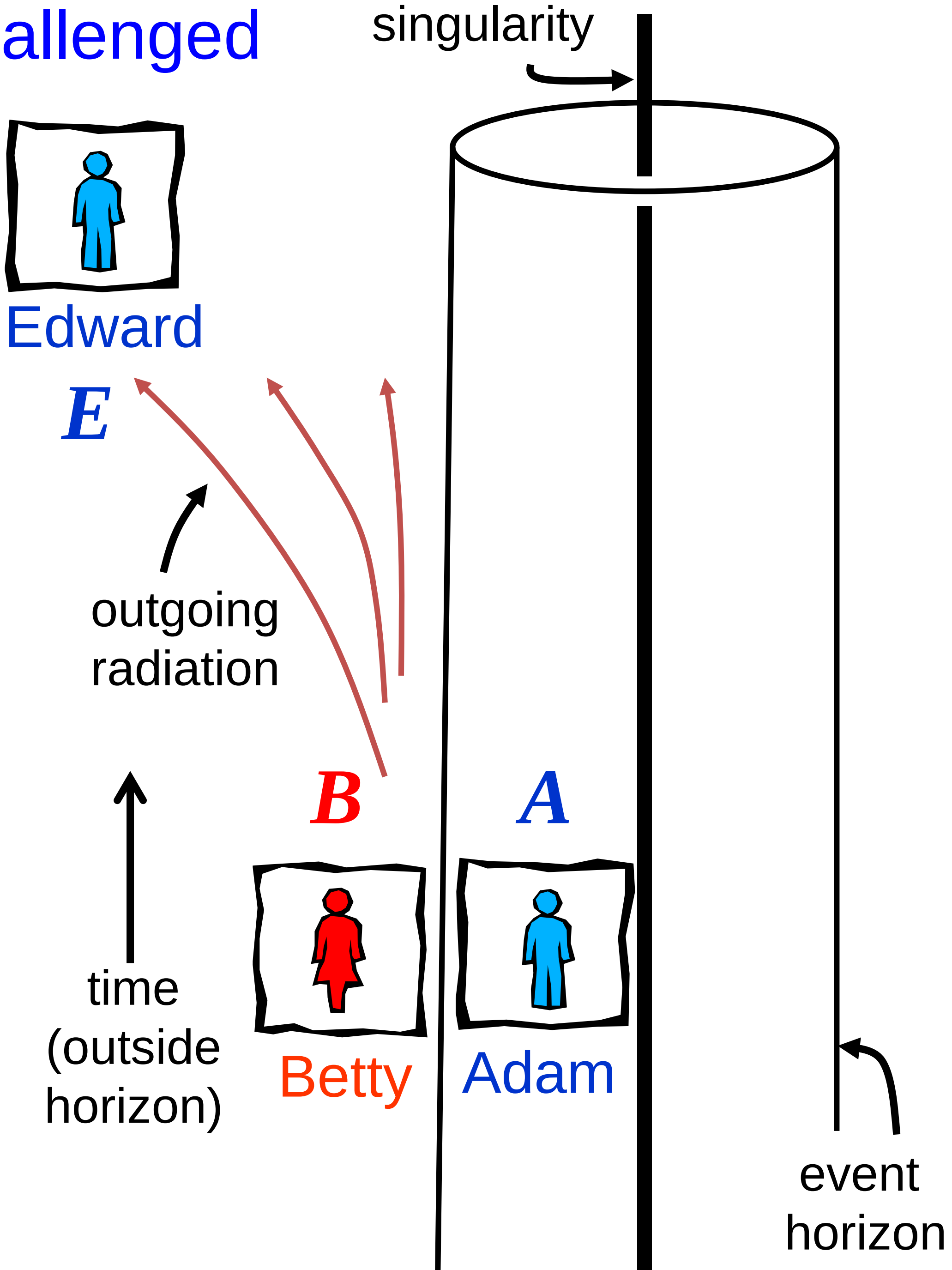
(AMPS 2012)

(1) For an old black hole, recently emitted radiation (*B*) is highly entangled with radiation emitted earlier (*E*) by the time it reaches Charlie.

(2) If *B* is entangled with *E* by the time it reaches Charlie, it was already entangled with *E* at the time of emission from the black hole.

(3) If freely falling observer sees vacuum at the horizon, then the recently emitted radiation (*B*) is highly entangled with modes behind the horizon (*A*).

Monogamy of entanglement violated!



Black hole complementarity reconsidered

Black hole complementarity was on the right track after all. For a black hole *H* that is entangled with its previously emitted radiation *E*, it may be correct to view at least part of the black hole interior as encoded in *E*.

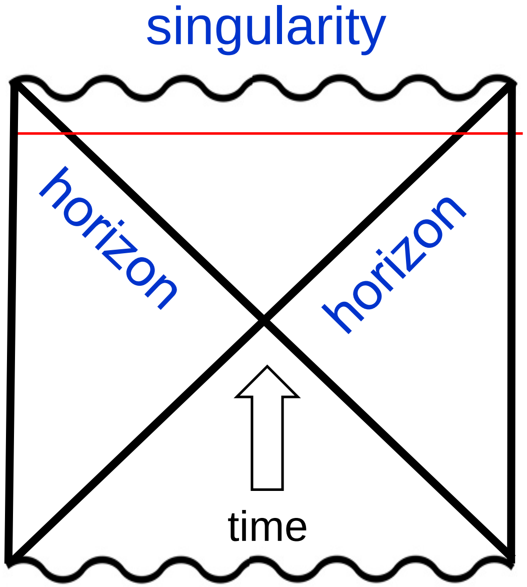
To reconcile complementarity with locality, we need some concepts that were not fully appreciated by AMPS: quantum error correction, computational complexity, and pseudorandomness.

The main idea: The encoding of the black hole interior is very robust, so that observers outside the black hole can disturb the entanglement between the interior and the recently emitted radiation only by performing operations on the early radiation *E* that are computationally infeasible. For realistic computationally bounded observers, complementarity and locality are compatible.

This is work with Isaac Kim and Eugene Tang. Our work builds on earlier results by a number of other people, including *Papadodimas and Raju 2012, Verlinde and Verlinde 2013, Harlow and Hayden 2013, Susskind and Maldacena 2013, Flammia, Haah, Kastoryano and Kim 2016, Ji, Liu and Song 2017*, …. It includes some new insights about quantum error correction against adversarial noise.

To explain this we need some further background.

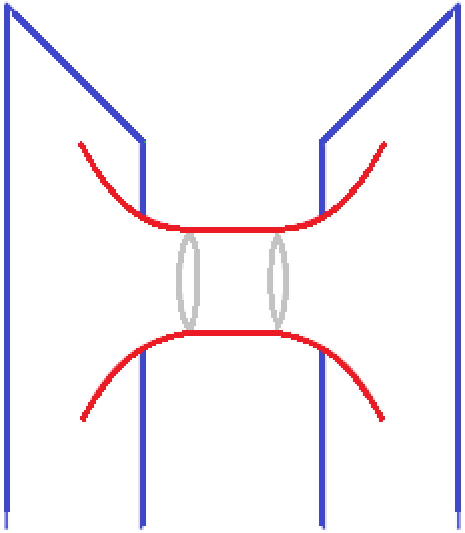
AMPS and the TFD



In the thermofield double (TFD) state, the AdS black hole on the right is maximally entangled with another system (the black hole on the left), yet its horizon is smooth. How do we reconcile with the AMPS argument?

The answer is clear in this case: Hawking radiation from the right BH can be maximally entangled with modes on the other side of the horizon, and with another system, because both are the *same* system (the left black hole).

We may boldly assert that this is the resolution in general: the interior geometry of the black hole is actually constructed from the system which is maximally entangled with the black hole.



We can imagine gravitationally collapsing the system entangled with the black hole, then applying a one-sided unitary to obtain the TFD. Identifying the radiation emitted long ago with the black hole interior seems wildly nonlocal, and just about audacious enough to be on the right track.

Classical Pseudorandomness

Two probability distributions on *n*-bit strings. Easy to distinguish in principle, but *hard to distinguish by any feasible computation*.

$$p_I(x) = \frac{1}{2^n}, \qquad \forall x \in \{0,1\}^n$$

$$p_S(x) = \begin{cases} 2^{-\alpha n}, & x \in S \\ 0, & x \notin S \end{cases}$$

Why? For a fixed circuit *C*, and a uniform distribution on all sets of bit strings with |S| = $2^{\alpha n}$ elements, probability that *C* "accepts" a sample is nearly the same for both distributions: p_I and p_S are distinguishable by more than an exponentially small amount with a probability that is doubly exponentially small.

$$\text{Prob}\left(|P_C(S) - P_C(I)| \geq \epsilon\right) \leq e^{-2|S|\epsilon^2}$$

And the number of circuits of size polynomial in *n* is much less than doubly exponentially large.

It may be computationally hard to sample from p_S, but there are other pseudorandom distributions that we can sample from efficiently (a highly plausible conjecture).

Quantum Pseudorandomness

Two quantum states of *n* qubits. Easy to distinguish in principle, but *hard to distinguish by any feasible quantum computation*.

$$\boxed{\rho_{\text{mixed}} = I/2^n} \qquad \boxed{\rho_{\text{pure}} = |\psi\rangle\langle\psi|}$$

Why? For a fixed observable *C*, with eigenvalues 0 and 1, and a uniform (Haar) distribution on all pure quantum states, probability that *C=1* (observable "accepts" state) is nearly the same for both states: ρ_{mixed} and ρ_{pure} are distinguished by more than an exponentially small amount with a probability that is doubly exponentially small.

$$\text{Prob}\left(\left|\langle\psi|C|\psi\rangle - \frac{\text{Tr}(C)}{2^n}\right| \geq \epsilon\right) \leq e^{-c2^n\epsilon^2}$$

And the number of quantum circuits of size polynomial in *n* is much less than doubly exponentially large.

It may be computationally hard to prepare the pure state *ψ*, but there are other pseudorandom quantum states that we can prepare efficiently (a highly plausible conjecture).

Quantum Pseudorandomness

The existence of *efficiently preparable pseudorandom quantum states* follows from a standard assumption of "post-quantum" cryptography: the existence of a family {PRF$_k$, *k* in *K}* of quantum-secure pseudorandom functions. Each PRF$_k$ is efficiently computable, but it's computationally hard to distinguish a randomly sampled PRF$_k$ from a random function. (*K* is the *key space* of the family.)

Example: *a uniform mixture of these pure states:*

$$|\phi_k\rangle = \frac{1}{\sqrt{N}} \sum_{x \in X} \omega_N^{\mathrm{PRF}_k(x)} |x\rangle, \quad \omega_N = e^{2\pi i/N}, \quad N = 2^n$$

Easy to prepare because PRF$_k$ is easy to compute. But the phases are computationally indistinguishable from random (*Ji, Liu, Song 2017*).

Since such states plausibly exist, and black holes are very effective scramblers of quantum information, *it is plausible that a partially evaporated black hole (or other strongly chaotic quantum systems) efficiently prepares Hawking radiation in a pseudorandom state after the Page time.* In that case, the key space is provided by black hole microstates. For a finite-temperature black hole, pseudorandom means information-theoretically but not computationally distinguishable from a perfectly *thermal* state.

Encoding the black hole interior

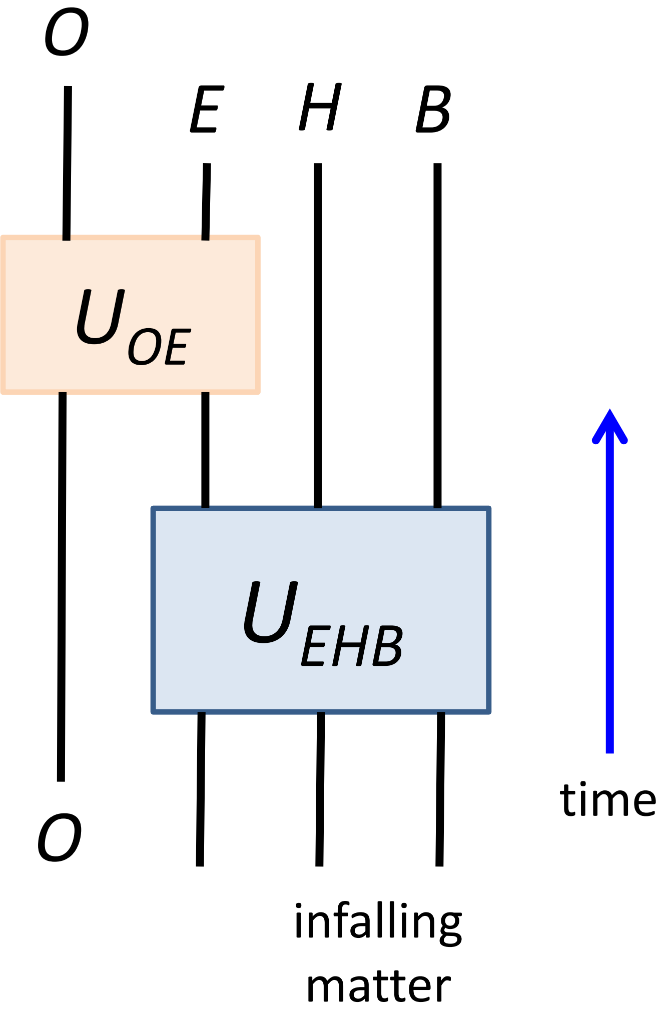
Old black hole, small observer:
$|O| << |H| << |E|$

E: early Hawking radiation
H: remaining black hole
B: recently emitted Hawking mode
O: observer

U_{EHB}: unitary black hole dynamics
U_{OE}: interaction of observer with radiation

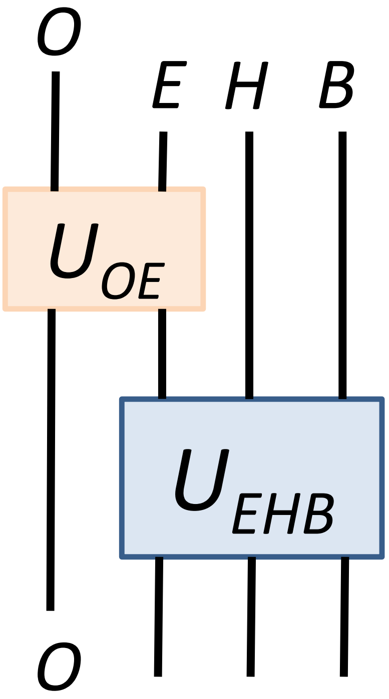
Idea: *B*'s partner mode *A* behind the black hole horizon is robustly encoded in *EH*.

Computationally bounded observer *O* cannot disrupt this encoding (cannot send an acausal signal to the black hole interior).



Encoding the black hole interior

Key assumption: Hawking radiation is pseudorandom. For an old black hole, where the size $|E|$ of the previously emitted radiation system is much larger than the size $|H|$ of the black hole, a computationally bounded (polynomial-time) observer O is unable to distinguish the state of E from a maximally mixed state with probability better than $\exp[-\alpha|H|]$.



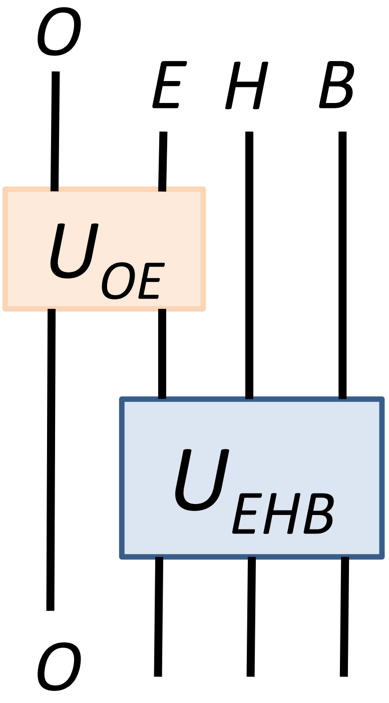
No two-outcome measurement with computationally complexity polynomial in |H| can distinguish state of EB from thermal state.

$$\left|\Pr\left(\mathcal{M}(\rho_{EB})=1\right)-\Pr\left(\mathcal{M}(\sigma_{EB}^{\max})=1\right)\right|\le 2^{-\alpha|H|}$$

Main result: If in addition $|O| << |H|$, then ghost logical operators can be constructed, which can be regarded as operators acting on the encoded black hole interior. These act on a subspace of EH which is (nearly) maximally entangled with the recently emitted radiation system B, and their action on the state of BEH (nearly) commutes with the action of any computationally bounded observer O acting on E.

Encoding the black hole interior

$$\boxed{\left|\Pr\left(\mathcal{M}(\rho_{EB})=1\right)-\Pr\left(\mathcal{M}(\sigma_{EB})=1\right)\right|\leq 2^{-\alpha|H|}}$$



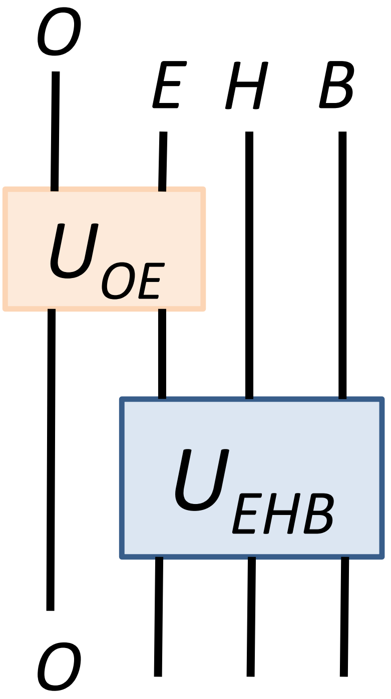
Main result: If in addition $|O| << |H|$, then ghost logical operators can be constructed, which can be regarded as operators acting on the encoded black hole interior. These act on a subspace of *EH* which is (nearly) maximally entangled with the recently emitted radiation system *B*, and their action on the state of *BEH* (nearly) commutes with the action of any computationally bounded observer *O* acting on *E*.

The argument has two steps. In the first step, we show that pseudorandomness of the Hawking radiation implies that the system *A* (entangled with *B*) encoded in the radiation is computationally hard to decode. This is a refinement of earlier arguments by *Harlow and Hayden 2013* and *Aaronson 2016*.

The second step is the construction of the ghost logical operators, logical operators of a quantum code that nearly commute with *O*'s actions.

Pseudorandomness and Decoupling

$$\left|\Pr\left(\mathcal{M}(\rho_{EB})=1\right)-\Pr\left(\mathcal{M}(\sigma_{EB})=1\right)\right|\le 2^{-\alpha|H|}$$



The late radiation *B* can be regarded as a "reference system" entangled with the encoded system embedded in *EH*. The observer may be regarded as an "environment" that "purifies" the noisy channel which afflicts the early radiation system *E*. We'll suppose |*O*| small compared to |*H*|, the size of the remaining black hole. (The "Kraus rank" of the channel is not too large.)

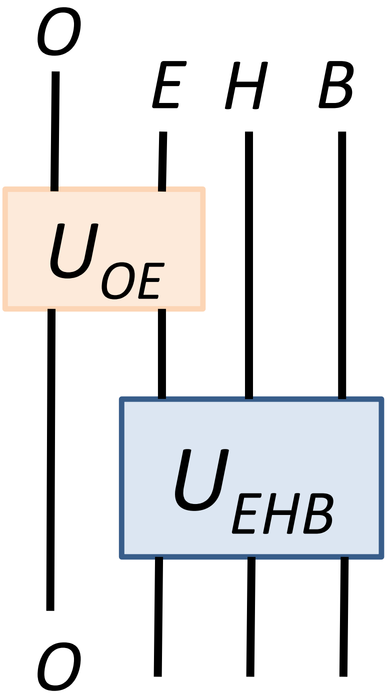
Then *O* and *B* nearly decouple:

$$\left\|\rho_{OB}-\rho_O\otimes\rho_B\right\|_1\le 6\cdot 2^{-(\alpha|H|-|O|)}$$

(The constant factor on the right-hand side actually depends on |*B*|; what's shown is the case where *B* is a single qubit.) This (approximate) decoupling condition means that the observer acquires negligible information about the encoded interior mode *A* that is entangled with the (late) exterior mode *B*.

Decoupling and Error Correction

$$\boxed{\| \rho_{OB} - \rho_O \otimes \rho_B \|_1 \le 6 \cdot 2^{-(\alpha|H|-|O|)}}$$



This decoupling condition holds because the observer performs an operation of polynomial complexity, and the Hawking radiation is assumed to be pseudorandom.

A consequence of decoupling is that the damage inflicted on the encoded system *A* (embedded in *EH*) that purifies *B* can be almost perfectly reversed by a suitable recovery map. (We are not making a claim about the complexity of the recovery map.) Recovery is possible because negligible information about the encode state leaks to *O*.

$$\max_{\rho} \| (\mathcal{R} \circ \mathcal{E})(\rho) - \rho \|_1 \le 2\sqrt{6}\, 2^{-(\alpha|H|-|O|)}$$

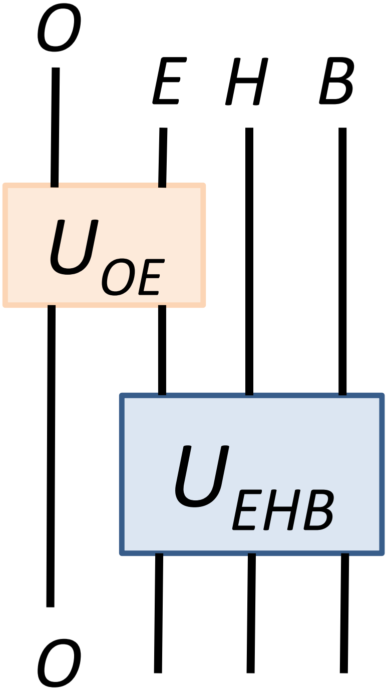
$\rho =$ encoded state, $\quad \mathcal{E} : E \to E =$ noisy channel, $\quad \mathcal{R} : EH \to EH =$ recovery map

Residual error after recover is exponentially small.

Decoupling and Error Correction

$$\boxed{\|\rho_{OB}-\rho_O\otimes\rho_B\|_1\le 6\cdot 2^{-(\alpha|H|-|O|)}}$$

$$\max_\rho \|(\mathcal{R}\circ\mathcal{E})(\rho)-\rho\|_1\le 2\sqrt{6}\,2^{-(\alpha|H|-|O|)}$$



That "post-quantum cryptography" assumptions imply the hardness of *decoding* Hawking radiation had been noted earlier by *Harlow and Hayden 2013, Aaronson 2016*.

We've related this hardness of decoding to pseudorandomness of the Hawking radiation. We have also seen that decoding remains hard even for |*H*| << |*E*|, provided |*H*| >> 1.

And we've seen that errors afflicted by noise with complexity polynomial in |*H*| is correctable. Usually we consider quantum error correction against noise that is weak and weakly correlated. In this case the noise can be strong and highly correlated, and chosen adversarially, as long as the noise process has reasonable computational complexity. The black hole encoding, or other encodings associated with pseudorandom states, can resist this correlated noise. (But note that the "noise" acts only on *E*; we treat *H* as a noiseless system.)

“Ghost Logical Operators”

We can construct logical operators for a quantum code which commute with correctable errors (when acting on the code space).

In the setting of *exact error correction*:

$$E_a\,|0\rangle_L \perp E_b\,|1\rangle_L, \quad E_a\,|+\rangle_L \perp E_b\,|-\rangle_L.$$

where E_a, E_b are correctable errors. Logical Pauli operators acting on code basis states,

$$\tilde{Z}\,|0\rangle_L = |0\rangle_L, \quad \tilde{Z}\,|1\rangle_L = -|1\rangle_L,$$
$$\tilde{X}\,|+\rangle_L = |+\rangle_L, \quad \tilde{X}\,|-\rangle_L = -|-\rangle_L,$$

can be extended:

$$\tilde{Z}E_a\,|0\rangle_L = E_a\,|0\rangle_L, \quad \tilde{Z}E_a\,|1\rangle_L = -E_a\,|1\rangle_L,$$
$$\tilde{X}E_a\,|+\rangle_L = E_a\,|+\rangle_L, \quad \tilde{X}E_a\,|-\rangle_L = -E_a\,|-\rangle_L,$$

These extended Pauli operators anticommute, square to 1, and *commute with correctable errors when acting on the code space*. The correctable errors do not disturb the logical operator algebra:

$$[\tilde{Z}, E_a]\Pi_{\text{code}} = 0 = [\tilde{X}, E_b]\Pi_{\text{code}}$$

Approximate Error Correction

The technical part of our work shows that a similar construction of ghost operators still works in the setting of *approximate error correction*.

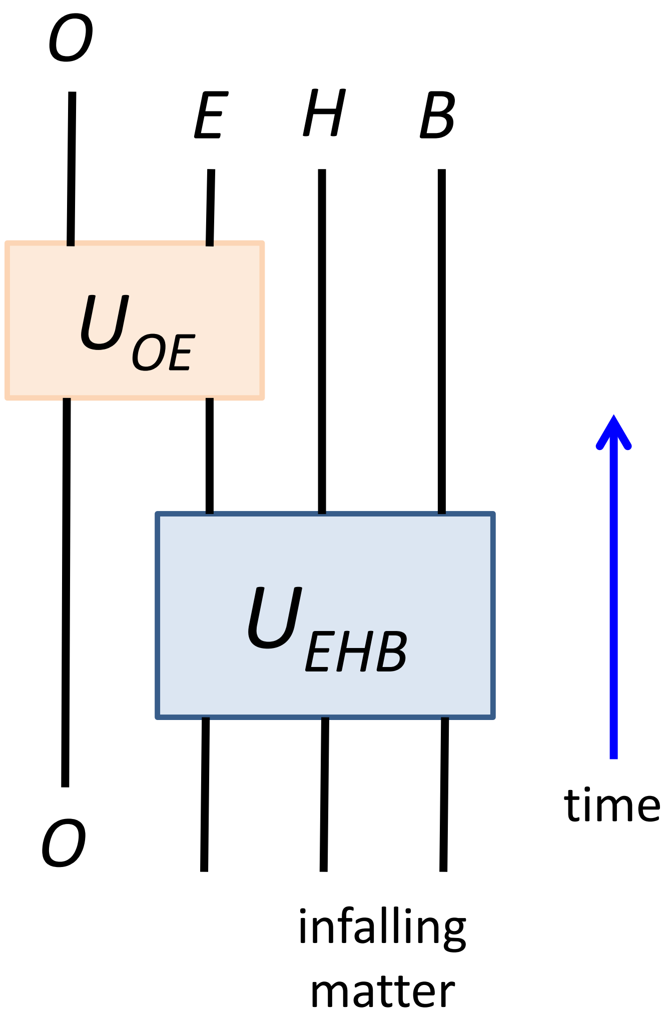
The (nearly) maximally entangled state of *B* with *EH* defines a (nearly) isometric map of the logical system *A* (the interior) to *EH*.

Our pseudorandomness assumption means that the computationally bounded observer O is (nearly) decoupled from *A* provided |O| << |H| (even if |H| << |E|).

This means that operations performed by observer are (nearly) correctable.

We can extend the action of logical operators to the space reached by correctable operators acting on the code space.

There are the "ghost operators" which (nearly) commute with the observer's operations.



Encoding the black hole interior

Plausibly, the Hawking radiation emitted by an incompletely evaporated old black hole is *pseudorandom*. (Hawking was *almost* right.)

The decoupling of computationally bounded observers from the reference system is quantified by the size $|H|$ of the remaining black hole, not the size $|E|$ of the radiation systems.

The Hawking radiation is hard to decode, as long as H is macroscopic.

It is hard to create a firewall by acting on the radiation -- causality is well respected from the viewpoint of computationally bounded observers. On the other hand, superpolynomially powerful quantum computes can tear spacetime apart.

Interior observers, who have access to the key space H as well as E can perform operations which are beyond the ability of the exterior observers.

The encoded interior is well protected against harsh actions by the exterior agents. For example, they can control probe systems which interact with all the radiation quanta outside the black hole without disturbing the encoded interior (at least until late in the evaporation process when $|H|$ is no longer very large).

Encoding the black hole interior

For effective field theory to be a good approximation, we require not only low curvature and low energy, but also *low complexity and low Kraus rank*.

We have shown that encodings with the desired properties exist, not that they really describe actions performed by observers who fall into the black hole.

The encoding of the black hole interior in *EH* is state dependent --- that is, the subspace of *EH* which is maximally entangled with the system *B* depends on what black hole microstate has been evaporating.

Is state dependence a problem? It suggests that we still don’t have a fully satisfactory understanding of measurements inside black holes.

Further research may clarify the connection between these robust encodings and the disconnected components of the entanglement wedge (“islands”) in AdS/CFT.

What’s the relation to complexity quantified by the size of the “python’s lunch”?

The general structure of this robust encoding of the black hole interior *may* apply beyond the context of AdS/CFT duality.

As we learn more, we may be better equipped to understand what quantum gravity tells us about very early universe cosmology.