ABSTRACT

Title of Dissertation: LOCALITY AND COMPLEXITY IN QUANTUM INFORMATION DYNAMICS

Dhruv Devulapalli
Doctor of Philosophy, 2025

Dissertation Directed by: Professor Andrew M. Childs
Department of Computer Science
Professor Alexey V. Gorshkov
Department of Physics

Locality constrains the flow of information between different parts of many-body quantum systems. In quantum computers, this affects the ability to perform arbitrary interactions for quantum information processing tasks. A crucial challenge for scalable quantum architectures is thus to minimize the overheads due to locality constraints. Locality constraints also affect the way information and entanglement can be spread in many body quantum systems, and our ability to make predictions about such systems.

One way to implement arbitrary interactions in locality-constrained architectures is by permuting qubits, also known as performing quantum routing. In the first part of this thesis, we investigate quantum routing strategies using emerging architectural capabilities beyond swap gates. We first investigate the use of mid-circuit measurements and classical feedback, which enable quantum teleportation. We prove upper bounds on the speedup afforded by the use of quantum teleportation for routing. We also design architectural connectivities which can exploit

teleportation-based speedups. We then consider quantum routing by continuous-time Hamiltonian evolution, designing a routing algorithm that achieves a polynomial speedup in routing time over gate-based routing methods in architectures with a bottleneck, which is an intermediate region that constrains interactions between the rest of the device. We improve the lower bound on the time taken for routing in such systems, and develop a matching upper bound on the closely related task of generating and distributing entanglement.

In the second part of this thesis, we investigate the behavior of finite-sized systems evolving by local Hamiltonians for a long duration. A major direction of research in condensed matter physics is the investigation of whether isolated systems thermalize (i.e., reach a state of thermal equilibrium). Various theoretical proposals such as information scrambling, entanglement between subsystems, and the eigenstate thermalization hypothesis have attempted to explain the conditions under which systems equilibrate. At the same time, notable counter-examples to thermalization have been found, such as systems that are integrable, exhibit many-body localization, or quantum many-body scars. We approach this question from a complexity-theoretic perspective. We show that the problem of deciding whether or not observables equilibrate to a given value is PSPACE-complete. We also show that deciding whether observables relax to the thermal equilibrium value is contained in PSPACE, and is PSPACE-hard under quantum polynomial time reductions. These problems are thus intractable even for a quantum computer. Our results indicate that the proposed conditions leading to thermalization are either difficult to decide, or are not generic indicators of thermalization.

LOCALITY AND COMPLEXITY IN QUANTUM INFORMATION DYNAMICS

by

Dhruv Devulapalli

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Advisory Committee:
Professor Andrew M. Childs, Chair/Advisor
Professor Alexey V. Gorshkov, Co-Advisor
Professor Alicia J. Kollár
Professor Nathan Schine
Professor Christopher Jarzynski, Dean's Representative

Dedication

To my grandparents

Acknowledgments

I am deeply grateful to my advisors Andrew and Alexey. Alexey has been a constant source of enthusiasm, new ideas, and support especially during times when I was not confident in my abilities as a researcher. Andrew has shown me how to approach and think about research problems, sharpened my writing, and inspired me with his patience and persistence in projects we have worked together on. They have both been generous with their time, supported my travel to conferences, and my growth as a researcher.

I would like to thank members of my committee, Christopher Jarzynski, Nathan Schine, and Alicia Kollár for their thoughtful questions and feedback.

I am grateful to all my collaborators: Eddie Schoute, Aniruddha Bapat, Andrew Guo, Connor Mooney, James Watson, Dominik Hangleiter, Zhenning Liu, Noah Berthusen, Michael Gullans, Daniel Gottesman, Yi-Kai Liu, Alicia Kollár, Nathan Constantinides, Ali Fahimniya, Trey Porto, Sharoon Austin, Abhinav Deshpande, Minh Tran, Chao Yin, Andy Lucas, Jeffrey Yu, and Alex Schuckert. I have truly enjoyed working with and learning from all of you.

I am grateful for the camaraderie and support of my friends and colleagues at QuICS and JQI. I learned a great deal from them, and enjoyed many research and non-research chats and time spent with them at conferences. In particular I'd like to thank Jeet Mohapatra, Elizabeth Bennewitz, Su-Kuan Chu, Emil Khabiboulline, Kunal Sharma, Chris Fechisin, Mahathi Vempati, Manasi Shingane, Matt Kovacs-Deak, Yifan Hong, Jacob Lin, Sean Muleady, Yuxin Wang,

Alex Schuckert, Zhi-Yuan Wei, Joe Iosue, Jeffery Yu, Twesh Upadhyaya, Joel Rajakumar, Dan Spencer, Srilekha Gandhari, Thomas Steckmann, Vikram Kashyap, Joe Carolan, Amin Shiraz Gilani, Daochen Wang, Shih-Han Hung, Jake Bringewatt, Adam Ehrenberg, Pradeep Niroula, Jim Garrison, Atul Mantri, Luke Schaeffer, Oles Shtanko, Hong Hao Fu, Luispe Garcia-Pintos, Lucas Brady, Przemek Bienias, Yaroslav Kharkov, Chris Baldwin, Sajjad Nezhadi, Samuel Tan, Jon Nelson.

I am grateful to Abhinav Deshpande and James Watson for teaching me much of what I know about complexity theory. I would also like to thank Kunal Sharma, Minh Tran, and Andrew Guo for their mentorship and advice. Special thanks to Yusuf Alnawakhtha for his feedback on thesis drafts, tactful writing tips, friendship over the years, and camaraderie in these last few months of my degree.

I am grateful to QuICS, JQI, and UMD staff members Andrea Svejda, Melissa Britton, Josiland Chambers for assistance with travel, finances, and everything administrative.

I would also like to thank my friends from undergraduate days who are also in grad school and have supported me and commiserated with me in the ups and downs of research. In particular, I am grateful to Karen Yu, Jyotirmai Singh, and Daniel Lengyel for many conversations about life, careers, and research.

Finally, my deepest gratitude and love to my wife Anjana, my parents, and my sister Prerna. Your love and support has made it possible for me to face difficulties and pursue my interests.

Table of Contents

Dedication ii

Acknowledgements iii

Table of Contents v

List of Figures vii

List of Abbreviations ix

Chapter 1: Introduction 1
- 1.1 Locality . . . 1
- 1.2 Quantum Computing . . . 2
- 1.3 Quantum routing and entanglement dynamics . . . 4
- 1.4 Long-time evolution and thermalization . . . 9
- 1.5 Outline . . . 13

Chapter 2: Quantum routing with teleportation 16
- 2.1 Introduction . . . 16
- 2.2 Preliminaries . . . 20
- 2.3 Bounds on routing time . . . 23
  - 2.3.1 Lower bounds . . . 23
  - 2.3.2 Upper bounds . . . 25
- 2.4 Faster permutations with teleportation . . . 26
- 2.5 Teleportation Advantage . . . 29
- 2.6 Bounding the Teleportation Advantage . . . 34
  - 2.6.1 Advantage for general graphs . . . 35
  - 2.6.2 Grids . . . 38
  - 2.6.3 Expander graphs . . . 39
  - 2.6.4 Hypercubes . . . 40
  - 2.6.5 Other graphs . . . 41
- 2.7 Discussion . . . 41

Chapter 3: Quantum routing and entanglement dynamics through bottlenecks 45
- 3.1 Introduction . . . 45
- 3.2 Preliminaries . . . 52
  - 3.2.1 Hamiltonian routing . . . 53

3.2.2 Gate-based routing . . . 54
3.2.3 Free-particle routing . . . 55
3.2.4 Transpositions . . . 58
3.2.5 Routing and entanglement distribution . . . 60
3.3 Routing free particles . . . 63
3.3.1 An optimal protocol for routing free particles . . . 63
3.3.2 Routing qubits . . . 65
3.3.3 Lower bound on the free-particle routing time . . . 67
3.4 Lower bounds on qubit routing through vertex bottlenecks . . . 72
3.5 Entanglement flow through bottlenecks . . . 81
3.6 Discussion . . . 87

Chapter 4: The complexity of thermalization in finite quantum systems 92
4.1 Introduction . . . 92
4.2 Preliminaries . . . 99
4.2.1 Algorithmic primitives . . . 100
4.2.2 Thermalization . . . 103
4.3 PSPACE-hardness of FSRelax . . . 108
4.3.1 Turing Machines and Setup . . . 109
4.3.2 TM-to-Hamiltonian Mapping . . . 110
4.3.3 Long-Time Expectation Values . . . 112
4.3.4 Putting it All Together and Verifying the Promise . . . 116
4.4 Containment of FSRelax in PSPACE . . . 118
4.5 FSTherm(MC) and FSTherm(Gibbs) are contained in PSPACE . . . 123
4.6 Hardness of FSTherm(MC) and FSTherm(Gibbs) . . . 135
4.6.1 Modifying the TM . . . 136
4.6.2 Modifying H, $\mathcal{H}$, and A . . . 138
4.6.3 A quantum polynomial time reduction from FSHalt to FSTherm(MC) . . . 140
4.6.4 A conjectured classical polynomial time reduction from FSHalt to FSTherm(Gibbs) 143
4.7 BQP-Hardness of FTFSRelax . . . 144
4.7.1 Modifying the TM-to-Hamiltonian Construction . . . 144
4.7.2 Time-Averaged Expectation Values . . . 145
4.8 Discussion . . . 147

Chapter 5: Outlook 151

Chapter A: Appendices for Chapter 2 156
A.1 Sparse routing . . . 156
A.2 Proof of diameter-expansion trade-off . . . 160

Appendix B: Appendices for Chapter 3 163
B.1 Proof of Lemma 3.12 . . . 163

Appendix C: Appendices for Chapter 4 177
C.1 Finite Time-Averaged Expectation Values . . . 177

List of Figures

1.1 (a) Circuit to create Bell pairs that does not respect connectivity of a 1D nearest-neighbour (1DNN) architecture. (b) Using a permutation to prepare the Bell pairs locally and respect 1DNN connectivity. . . . 5
1.2 Quantum repeater protocol enabling teleportation across arbitrary distances in constant depth. . . . 7
1.3 Star graph S_9. . . . 9

2.1 Constant-depth long-range teleportation protocol on a path of 7 qubits. The X and Z gates are classically controlled by the parities of the two sets of measurement results. This protocol can be extended to paths of any length without increasing the circuit depth. . . . 18
2.2 (a) A grid architecture with ancillas (blue) interspersed between data qubits (black). The red ovals indicate which ancilla corresponds to each data qubit. (b) An equivalent architecture in our model. . . . 21
2.3 Permutations on a 1D lattice . . . 27
2.4 Permutation π^l_{wheel} (shown by red double-sided arrows) that exchanges l pairs of vertices on the wheel graph W_{N+1} (shown in black). . . . 28
2.5 The graph $L(n)$. The black lines show edges between layers, while blue lines show edges within a layer (colored for visibility). . . . 31

3.1 The star graph on 9 vertices, S_9. . . . 47
3.2 A tripartition with vertex bottleneck C. Parts L, C, R contain N_L, N_C, N_R qubits, respectively. By the tripartite connectivity constraint, there are no edges between qubits in L and qubits in R. . . . 48
3.3 A global transposition permutation on the star graph. The black lines denote the connectivity constraints. The red lines with arrows indicate pairs of qubits to be swapped. . . . 59
3.4 Any cyclic permutation can be decomposed into two stages of transpositions. For example, a clockwise cycle of all vertices can be achieved by a global transposition between pairs of sites connected by blue edges followed by a global transposition for red edges. To perform an anti-clockwise cycle, simply transpose along the red edges and then the blue edges. . . . 60
3.5 Routing enables entanglement distribution. (a) The system has two parts L, R that are initially entangled with hidden auxiliary systems l, r, respectively. (b) On swapping every qubit in L with the corresponding qubit in R, the L system becomes entangled with r, and R becomes entangled with l, so we have generated 6 ebits of entanglement across the cut dividing L and R. . . . 61

3.6 Partitioning of system for SIE. 62
3.7 Swapping the k^{th} Fourier modes on the left and right sets of leaves. 65
3.8 Linear entangling rate between N leaves of the star graph and the center vertex. If the initial state is a product state of the N-qubit GHZ state $\frac{1}{\sqrt{2}}(|00\ldots0\rangle + |11\ldots1\rangle)$ on the leaves, and $|0\rangle$ on the center, then evolving for time π/N with interaction $h = |1-\rangle\langle 1-|_{l_i,c}$ (i.e., the generator of a CNOT) on each edge results in a final $(N+1)$-qubit GHZ state on all the vertices. 83
3.9 The vertex barbell graph with $2N+1$ vertices consists of two copies of the complete graph K_N connected by a common vertex. Here, we depict the vertex barbell with 11 vertices. 89

4.1 The relaxation decision problem. For a given initial state $|\psi\rangle$, we plot the time-average of $\mathcal{A}_N := \frac{1}{T}\int_0^T dt \langle\psi_0|e^{iHt}\mathcal{A}_N e^{-iHt}|\psi_0\rangle$. A^* is indicated by the dashed green line. Given the promise that $\bar{\mathcal{A}}_N := \lim_{T\to\infty}\langle\mathcal{A}_N\rangle_T$ is such that either $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$ or $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$, FSRelax is the problem of deciding which of these two holds. 95
4.2 The solid green and dashed-green lines represent the time-averaged expectation values in the halting and non-halting cases respectively, as p is tuned. The red line represents the expectation values with respect to the microcanonical ensemble as p is tuned. 140

A.1 Advancing a train of tokens, as in Algorithm A.1.1. Blank vertices hold state $|0\rangle$. The dashed lines represent connections with the local ancilla qubits. 156
A.2 Joining trains of tokens. 157

List of Abbreviations

GHZ	Greenberger-Horne-Zeilinger
ETH	Eigenstate Thermalization Hypothesis
STE	Small Total Entangling
SIE	Small Incremental Entangling
1DNN	1-Dimensional Nearest Neighbors

Chapter 1: Introduction

1.1 Locality

Locality is the principle that any object can only directly influence its immediate surroundings. For distant objects to interact, by the principle of locality, there must be some kind of medium between them through which the interaction can propagate. Locality is ubiquitous in physics and is fundamental in theories of classical mechanics, electromagnetism, and special and general relativity, where the maximum speed at which causal influence can travel is the speed of light. Quantum mechanics, however, challenges our notion of locality, as quantum theory violates Bell inequalities and therefore is incompatible with 'local realism' [NC10]. Nevertheless, quantum mechanics retains a notion of local causality through the no-communication theorem [PT04], which states that information cannot be transferred between spatially separated observers faster than the speed of light. In many-body quantum systems, locality manifests as a constraint on interactions between different parts of the system. Information spreads causally only by the means of local interactions between constituent subsystems, and is therefore constrained by *emergent* speed limits [ALY23]. This fact affects the design and behavior of quantum computers.

1.2 Quantum Computing

Applications of quantum computing and quantum information processing offer great promise, with possible implications for discovering new drugs, probing chemical reaction dynamics, designing better industrial catalysts, solving algebraic problems efficiently, high-precision measurements, and simulating physical systems [Pre18; Bay+21]. The wide range of these promises has spurred interest in the development of large-scale fault tolerant quantum computers. A crucial aspect of scalable quantum architectures is the fact that they are constrained by locality. Even in the very early days of quantum computing, it was suggested that parts of the device would be locally interconnected, motivated by the inherent difficulty of implementing arbitrary interconnections. [Fey82]. Every major modern proposal for a scalable quantum architecture faces connectivity constraints. In super-conducting quantum architectures, it is challenging to engineer highly connected qubits and non-planar architectures due to issues like crosstalk, interconnect crowding, and couplers crossing [Kja+20; LS25]. In ion trap quantum computers, although ions within a trap are all-to-all connected, this connectivity is not scalable, and increased numbers of ions in a trap lead to increased gate times. Proposals for scalable ion trap quantum computing make use of modular architectures of multiple ion traps connected via photonic interfaces [MK13] or shuttle ions between interaction zones or modules [Ami+10; Mos+23; Kau+20; Pin+21]. Limitations of connecting these modules therefore impose locality constraints on ion trap quantum computers as well. State of the art proposals for photonic quantum computers route photons from resource state generators to photon detectors, 'fusion devices' (that perform entangling measurements) and optical elements such as beam splitters [Bar+23; Bom+21; Bou+21]. Furthermore, the resource state generators are non-deterministic and must be multiplexed to generate sufficient states

on demand for computation. Modular approaches to photonic quantum computers connect modules via photonic interconnects [LN22]. Routing photons, connecting modules, and multiplexing resource state generators hence introduce a notion of locality constraints for photonic quantum computers. Even neutral atom quantum computers, which have gained much attention lately due to their powerful reconfigurability [Blu+22; Blu+24], do not have simultaneous all-to-all connectivity between every pair of qubits [Tan+24a]. Even at the level of logical qubits, we expect large numbers of redundant physical qubits to be used to create each logical qubit, and the interaction between these logical qubits will be constrained by the interactions between different groups of corresponding physical qubits.

While scalable quantum architectures are connectivity-constrained, quantum information processing tasks often make use of interactions between arbitrary pairs of qubits. In order to implement interactions between distant qubits, we must make use of interactions between intermediate qubits to move quantum information around. Due to the finite speed of propagating quantum information [LR72], we thus incur time overheads to implement arbitrary circuits that do not respect the connectivity constraints.

In addition to inducing an overhead in the implementation of quantum algorithms, locality constraints affect the way entanglement can be generated and spread in many body quantum systems. Aside from quantum information processing applications, the fact that information takes time to spread throughout a quantum system also affects whether the system behaves like a statistical ensemble. When information that was initially local has propagated through the entire system, correlations become highly non-local. When local information becomes delocalized, and local subsystems no longer retain a memory of the initial microstate (i.e., the information is not accessible to local probes), the system is often said to have become scrambled [Swi18]. We

can thus make predictions about observables in such systems using statistical ensembles, entirely based on macroscopic quantities of our initial state such as number of particles, energy, temperature. The process by which a system equilibriates and can be described by such ensembles is called *thermalization*, and is a central topic of study in modern condensed matter physics.

Motivated by applications to scalable quantum computing, and behavior of many-body quantum systems, this thesis studies the following two questions:

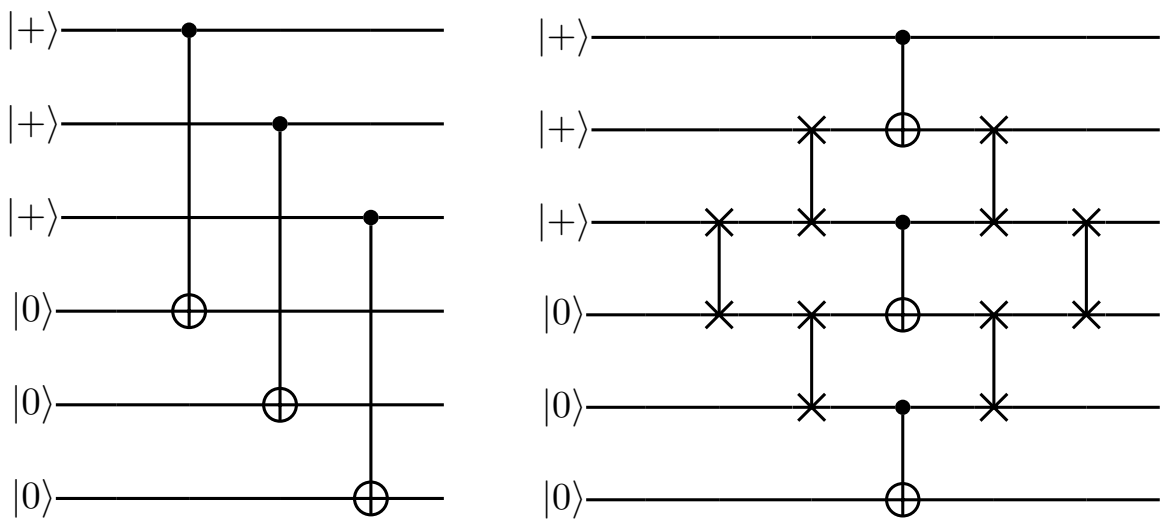
1. How fast can we engineer flow of information in quantum systems?

2. How fast does information flow in quantum systems when we do not control it, and how hard is it to make predictions about such uncontrolled systems?

We study these questions through the lens of quantum routing, entanglement dynamics, and thermalization, applying tools from computational complexity theory. We now give an overview of these topics, and outline the contributions of this dissertation.

1.3 Quantum routing and entanglement dynamics

A natural approach to implementing long-range interactions in the presence of connectivity constraints is by moving quantum information around using local interactions. Moving the state of a single qubit is known as quantum state transfer [Bos03; Bos07]. State transfer can be used to implement interactions between arbitrary pairs of qubits, by simply transferring their states to an adjacent pair of qubits. The speed at which quantum information can be transferred is limited by Lieb-Robinson bounds [LR72; ALY23], which therefore lower bound the overhead of implementing arbitrary quantum circuits.



(a) (b)

Figure 1.1: (a) Circuit to create Bell pairs that does not respect connectivity of a 1D nearest-neighbour (1DNN) architecture. (b) Using a permutation to prepare the Bell pairs locally and respect 1DNN connectivity.

Quantum routing generalizes state transfer to moving multiple qubits simultaneously. More formally, quantum routing is the task of permuting an unknown input state. On an architecture described by a graph G with vertices V and edges E, routing aims to implement arbitrary permutations $\pi : V \to V$ such that for any product state $|\psi\rangle = |\psi_1\rangle|\psi_2\rangle \dots |\psi_{|V|}\rangle$ we obtain $|\psi\rangle = |\psi_{\pi^{-1}(1)}\rangle|\psi_{\pi^{-1}(2)}\rangle \dots |\psi_{\pi^{-1}(|V|)}\rangle$ after routing. Furthermore, the permutation must be implemented only using interactions that obey the connectivity constraint dictated by the edges E. By performing permutations, we can perform multiple state transfers in parallel, reducing the overhead of implementing arbitrary circuits in a connectivity-constrained architecture.

One way to perform routing is by making use of swap gates, since any permutation can be decomposed into a product of local swaps. An example of this is depicted in Fig. 1.1a and Fig. 1.1b, where we show how a circuit that does not respect 1DNN connectivity can be trans-

formed to one which does by the use of swap gates. The problem of performing routing using swap gates has been well-studied in the context of classical networks [ACG94]. In quantum systems, however, we have access to more general interactions than swap gates. For example, a crucial ingredient for scalable fault tolerant quantum computers is circuit adaptivity, or the ability to apply gates conditioned on mid-circuit measurement outcomes [Ter15]. This enables correcting errors on their detection, and more broadly enables the paradigm of measurement-based quantum computing [Bri+09]. In recent years the ability to perform mid-circuit measurements and adaptively apply gates has become more readily available experimentally. In addition to its applications in error correction, this has prompted theoretical and experimental investigations into the use of adaptivity to prepare quantum states or implement important computational primitives in low depth [Bäu+24; PS13; BSW25; BW25; PSC24; Tan+24b]. Typically, propagating (classical) information between parts of a device using a classical controller is much faster than by using native gates on the quantum hardware. By performing a mid-circuit measurement in one part of the device, and conditionally applying a gate elsewhere, we can therefore transfer quantum information much faster than the speed limit obtained from Lieb-Robinson bounds. For example, mid-circuit measurements and feedback enable quantum teleportation across arbitrary distances in constant depth, as in Fig. 1.2, which suggests its application to quantum routing. In Chapter 2, we investigate the use of teleportation for routing and its strengths relative to swap-based routing methods. We find that teleportation obtains a speedup over swap-based methods for broad classes of permutations. In some special cases, teleportation can provide a speedup even for permutations in which qubits are routed short distances to destinations near their origin. We also obtain bounds on the maximum speedup teleportation obtains over swap-based methods for the worst-case routing time.

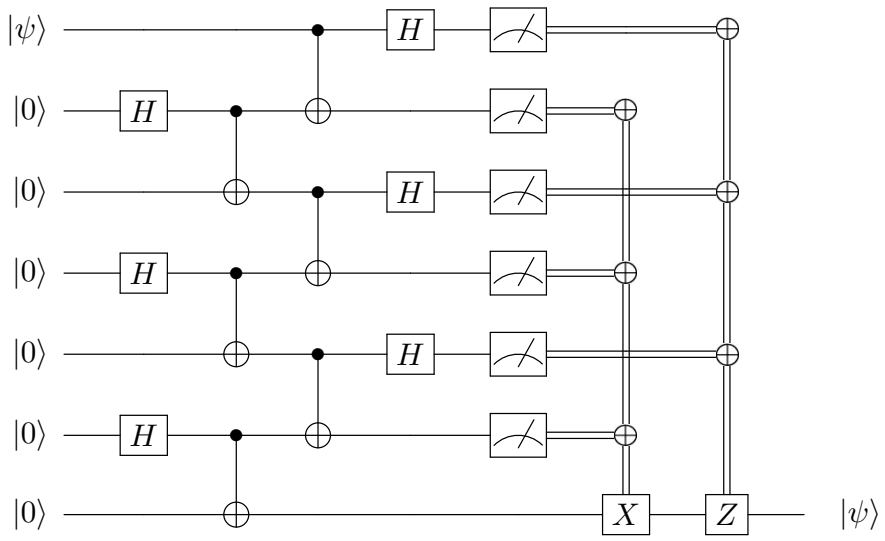


Figure 1.2: Quantum repeater protocol enabling teleportation across arbitrary distances in constant depth.

A closely related task to routing is distributing and generating entanglement. The rate at which entanglement can be generated between parts of a system constrains our ability to perform many quantum information tasks, as entanglement is a critical resource for tasks such as quantum algorithms [Joz98; JL03; EJ98; BM11; ZZC25], quantum error correction [Bra+25; Ben+96a; KL97], quantum communication [BCD01; GT07; Hor+09], and quantum sensing [DRC17; GLM04; Bol+96]. The fundamental importance of entanglement and its application to a wide range of quantum information tasks has prompted a large number of investigations into entanglement dynamics [VMV13; Mar+16; CVC05; Cub+03; Bra07; Dür+01; Ben+03]. Since routing can be used to distribute entanglement, bounds on entanglement dynamics can therefore provide bounds on the time taken to perform routing as well.

Previous results on entanglement dynamics have shown that the ability to generate entan-

glement between parts of a system is constrained by *bottlenecks*, or intermediate parts of the system with limited size. For example, the Small Incremental Entangling theorem (SIE) [VMV13] implies that the rate at which entanglement can be generated between two parts of a bipartition of a system is constrained by the number of edges joining them. Therefore, when the number of edges joining two large sets of vertices in an architecture is small, it takes a long time to generate a volume-law amount of entanglement between the two sets (and hence to perform routing). Such a set of edges is known as an *edge bottleneck*. For example, in a d-dimensional array of side length N with a total of N^d vertices, the number of edges joining two halves of the system scales as $\Theta(N^{d-1})$, which is asymptotically smaller than the number of vertices on each side ($\Theta(N^d)$), and hence these edges form a bottleneck.

Similarly, for a given bipartition of a graph, the vertex boundary is the set of vertices in each partition that are adjacent to the other partition. The Small Total Entangling property (STE) [Mar+16] states that the maximum amount of entanglement that can be generated by a unitary acting on the vertex boundary is bounded by twice the minimum dimension of the boundary of either partition. Hence, when a large number of vertices in each partition are only connected through a small vertex boundary, it takes a long time to perform routing across the partition. A vertex boundary that is much smaller than the size of the partitions is thus called a *vertex bottleneck*. A characteristic example of a vertex bottleneck is the star graph S_N (Fig. 1.3), which is the complete bipartite graph $K_{1,N-1}$ between one (center) vertex and $N-1$ leaf vertices.

The lower bounds on routing that can be obtained from STE are found to be tighter than the bounds from SIE [Bap+23]. While both SIE and STE apply to routing by a circuit of 2-qubit gates, or *gate-based* quantum routing, only SIE applies to routing by a continuous time Hamiltonian evolution, or *Hamiltonian* quantum routing. This gap between the previously known

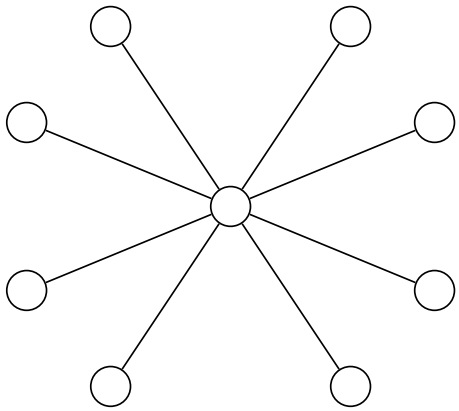


Figure 1.3: Star graph S_9.

lower-bounds suggests that speedups may be possible for Hamiltonian routing over gate-based methods. In Chapter 3, we investigate speedups for Hamiltonian routing over the gate-based model in systems with vertex bottlenecks. We improve the lower bound on Hamiltonian routing, proving a bound of $\Omega(\sqrt{N})$ on the Hamiltonian routing time of the star graph, in contrast to the previously best known bound of $\Omega(1)$. We generalize this result to vertex bottlenecks beyond the star graph. We also design a protocol that saturates this bound in systems of free particles. We further investigate the entanglement dynamics in systems with a vertex bottleneck, and develop a notion by which we obtain tighter bounds on the entanglement dynamics of continuous evolution. We describe these results in more detail in Section 1.5.

1.4 Long-time evolution and thermalization

Statistical mechanics has been remarkably successful at describing the physics of systems of large numbers of particles. In statistical physics, we model these large systems in terms of ensembles based on their macroscopic properties, such as number of particles, temperature, volume [PB21]. This approach allows us to drastically reduce the number of relevant degrees of freedom.

It is perhaps surprising that such an approach yields accurate descriptions of behaviour of real systems found in the lab and in nature. After all, a given set of macroscopic observables generally corresponds to a very large number of possible microstates, or specific, fully determined states. Nevertheless, we usually find that despite starting out in a highly specific microstate, systems tend to evolve in such a way that after sufficiently long time, their behaviour is well described by statistical ensembles based on their macroscopic properties. The process by which this takes place is known as thermalization. For example, let us consider an isolated system with Hamiltonian H starting in an initial state which has an expected energy E. If the system thermalizes, then it is found to be well described after sufficiently long time by the quantum microcanonical ensemble $\rho_{MC}(H, E)$, defined as

$$\rho_{MC}(H, E) = \frac{1}{W} \sum_i f\left(\frac{\lambda_i - E}{w}\right) |\psi_i\rangle\langle\psi_i|, \tag{1.1}$$

where f is an energy window function (typically decaying rapidly in its argument, so that $\rho_{MC}(H, E)$ is only supported on states with energy $\approx E$), $W = \sum_i f\left(\frac{\lambda_i - E}{w}\right)$ and w is a characteristic energy width. In systems that thermalize, at sufficiently long times, regardless of the details of the initial state other than E, observable expectations agree with their values for $\rho_{MC}(H, E)$.

The exact mechanism by which quantum systems thermalize is still not well understood. Classically, thermalization is hypothesized to occur due to ergodicity in chaotic systems. Since chaotic systems are ergodic, after a long enough time they have evolved uniformly through their phase space (constrained to a manifold specified by macroscopic variables determined by the initial state), and we can thus treat long-time observables using ensembles based on uniform mixtures of the phase space. However, this explanation does not transfer easily to quantum

systems. Isolated quantum systems evolve unitarily, and therefore do not sample the entire phase space. One proposed explanation for how isolated quantum systems thermalize is the eigenstate thermalization hypothesis (ETH) [Sre94; Deu91; Deu18]. According to the ETH, observable expectations for *every* eigenstate agree with their expectations for the thermal state. ETH is fulfilled for observables with diagonal elements (in the energy eigenbasis) that vary smoothly as a function of energy, and with small off-diagonal elements, as such observables evolve over long times with an expectation value agreeing with the microcanonical ensemble. While the ETH has been successful in explaining thermalization in a wide variety of systems [GE16], there are systems that thermalize without relying on ETH [GE16], and indeed where ETH explicitly does not hold [HH23; RS12]. There are also systems which equilibriate, but not to a thermal state, and retain memory of the initial state [GME11].

One proposed explanation for thermalization in quantum systems is information scrambling and entanglement between subsystems. When local information is spread, states that differed on local observables become difficult to distinguish as the local information is difficult to retrieve [Swi18]. This can be illustrated in terms of entanglement as well. Consider a system of spins with a single spin entangled with some reference system. At long times, the entanglement with the reference system is spread over the whole system (known as scrambling), and to recover it requires access to a large number of spins. In chaotic spin chains, information spreads ballistically at the Lieb-Robinson velocity, which suggests an explanation for thermalization in terms of scrambling. Thermalization and equilibration have also been linked to large entanglement in the energy eigenbasis. Entanglement has been suggested as the primary cause of equilibration, removing the need for the common postulate that systems can be considered a priori to be equiprobable over states with a given energy [PSW06]. This result, which is kinematic rather

than dynamical, argues that since almost all eigenstates of a Hamiltonian acting on a system and a large bath are highly entangled, any local (reduced) state of the system is close to the local state of an ensemble.

Another major area of research in quantum statistical physics is systems which *do not* thermalize, such as systems that exhibit many-body localization [Aba+19], quantum many-body scars [Tur+18], and integrable systems [Deu18; KWW06] (although integrable systems can equilibrate to a *Generalized Gibbs Ensemble*, which takes into account the extensive conserved quantities [Rig+07; GE16]). Understanding when and why thermalization occurs has driven a large volume of theoretical and experimental research. A related, and underlying question to this is whether we can predict long-time average expectations of observables. In addition, although there have been many efforts towards understanding the causes of thermalization, our understanding about the timescales required for relaxation and thermalization remains limited [Tas16; GE16]. In Chapter 4, we approach these questions from a complexity theoretic perspective, showing that they are computationally intractable. Our work shows that the thermalization process can encode a difficult computation, namely, a PSPACE-complete computation, even in systems that obey 1D local connectivity and have translationally invariant Hamiltonians. We also investigate the problem of estimating long-time observables, showing that it is PSPACE-complete. Our work builds off of previous results [SM21] showing that these problems are undecidable in the thermodynamic limit. Our work is motivated by the fact that real systems encountered in nature and the lab are finite-sized, and shows that the hardness is inherent in the thermalization process and not due to specifics of infinite-sized systems. A consequence of our hardness proofs is that there is no generic, easy way to determine whether a system thermalizes. Hence, our results imply that conditions such as the ETH and information scrambling are either easily decidable, or they cannot

be generic indicators of thermalization.

1.5 Outline

In Chapter 2, we investigate quantum routing with mid-circuit measurements and classical feedback. In particular, we show examples of speedups over swap-based and more general unitary routing methods by distributing entanglement and using LOCC to perform quantum teleportation. We further describe an example of an interaction graph for which teleportation gives a logarithmic speedup in the worst-case routing time over swap-based routing. We also study limits on the speedup afforded by quantum teleportation—showing an $O(\sqrt{N \log N})$ upper bound on the separation in routing time for any interaction graph—and give tighter bounds for some common classes of graphs.

In Chapter 3, we consider the entanglement dynamics and routing between two regions only connected through an intermediate "bottleneck" region with few qubits. We improve the lower bound on the routing time in systems with a vertex bottleneck. Specifically, for any system with two regions L, R with N_L, N_R qubits respectively (and $N_L \geq N_R$), coupled only through an intermediate region C with N_C qubits, for any $\delta > 0$ we show a lower bound of $\Omega(N_R^{1-\delta}/\sqrt{N_L}N_C)$ on the Hamiltonian quantum routing time when using piecewise time-independent Hamiltonians, or time-dependent Hamiltonians subject to a smoothness condition. We also prove an upper bound on the average amount of bipartite entanglement between L and C, R that can be generated in time t by such architecture-respecting Hamiltonians in systems constrained by vertex bottlenecks, improving the scaling in the system size from $O(N_L t)$ to $O(\sqrt{N_L}t)$. As a special case, when applied to the star graph (i.e., one vertex connected to N leaves), we obtain an $\Omega(\sqrt{N^{1-\delta}})$

lower bound on the routing time and on the time to prepare $N/2$ Bell pairs between the vertices. We also show that, in systems of free particles, we can route optimally on the star graph in time $\Theta(\sqrt{N})$ using Hamiltonian quantum routing, obtaining a speed-up over gate-based routing, which takes time $\Theta(N)$.

In Chapter 4, we investigate the problem of deciding whether a finite-sized system thermalizes or relaxes to a given value. Using tools from computational complexity theory, we show that deciding whether an observable relaxes to a given value is PSPACE-complete, and hence intractable even for a quantum computer. We also show that deciding whether a system thermalizes is PSPACE-hard, and with some assumptions, contained in PSPACE under quantum polynomial time reductions.

Citations to previously published work and preprints

Most of the work appearing in this thesis is either published or has appeared on the preprint server arXiv. We mention these here.

- Chapter 2: "Quantum routing with Teleportation", D. Devulapalli, E. Schoute, A. Bapat, A. M. Childs, and A. V. Gorshkov, Phys. Rev. Res. 6, 033313 (2024).

  Further work related to routing with teleportation appeared in

  - "Toward a 2D Local Implementation of Quantum Low-Density Parity-Check Codes", N. Berthusen, D. Devulapalli, E. Schoute, A. M. Childs, M. J. Gullans, A. V. Gorshkov, and D. Gottesman, PRX Quantum 6, 010306 (2025).

- Chapter 3: "Quantum Routing and Entanglement Dynamics Through Bottlenecks", D. De-

vulapalli*, C. Yin*, A. Y. Guo, E. Schoute, A. M. Childs, A. V. Gorshkov, and A. Lucas, arXiv:2505.16948. (*co-first-authors)

Further work related to routing and unitary synthesis appeared in

- "Optimal Routing Protocols for Reconfigurable Atom Arrays", N. Constantinides, A. Fahimniya, D. Devulapalli, D. Bluvstein, M. J. Gullans, J. V. Porto, A. M. Childs, and A. V. Gorshkov, arXiv:2411.05061.
- "Implementing a fast unbounded quantum fanout gate using power-law interactions", A. Y. Guo, A. Deshpande, S.-K. Chu, Z. Eldredge, P. Bienias, D. Devulapalli, Y. Su, A. M. Childs, and A. V. Gorshkov, Phys. Rev. Res. 4, L042016 (2022).

- Chapter 4: "Complexity of thermalization in finite quantum systems", D. Devulapalli, T.C. Mooney, J.D. Watson, arXiv:2507.00405.

Further work related to complexity theory and quantum computational advantage appeared in

- "Efficiently Verifiable Quantum Advantage on Near-Term Analog Quantum Simulators", Z. Liu, D. Devulapalli, D. Hangleiter, Y.-K. Liu, A. J. Kollár, A. V. Gorshkov, and A. M. Childs, PRX Quantum 6, 010341 (2025).

Chapter 2: Quantum routing with teleportation

2.1 Introduction

Common theoretical models of quantum computation assume that 2-qubit gates can be performed between arbitrary pairs of qubits. However, in practice, scalable quantum architectures have qubit connectivity constraints [Aru+19; MK13], which forbid long-range gates. These connectivity constraints are typically represented by a simple graph, where vertices correspond to qubits, and edges indicate pairs of qubits that can undergo 2-qubit gates. A quantum architecture with N qubits is thus represented by a graph G with N vertices. Circuits that use all-to-all connectivity must be transformed to new circuits that respect the architecture constraints specified by this graph. Simple transformations introduce polynomial overhead in the worst case, so it is crucial to lower this overhead.

A natural approach to mapping circuits to respect interaction constraints is by permuting qubits using *routing* protocols. Routing refers to the task of permuting packets of information, or *tokens*, on vertices of a graph. In *quantum* routing, tokens are data qubits, to be permuted on the graph specified by the architecture's connectivity constraints. Previous work has used swap gates to perform routing [Cow+19; CSU19], and routing protocols from a classical setting using swap gates [ACG94; Chu96; Zha99] can be naturally applied to the problem of routing quantum data as well.

Faster routing protocols can be obtained by using a wider range of quantum operations. For example, Hamiltonian evolution can obtain a constant-factor speedup over swap-based routing [Bap+21]. More details of the comparisons and advantages of quantum routing models to classical routing models can be found in reference [Bap+23]. However, these approaches rely on locality-restricted unitary evolution, so the routing time is limited by the propagation speed of quantum information [LR72; Bap+23].

In this work, we additionally allow for fast local operations, measurement and feedback (LOCC). Since this model allows for fast classical communication across long distances, it is not similarly constrained by the propagation speed of quantum information. For example, without prior shared entanglement, quantum teleportation over arbitrary distances can be performed in constant depth by using entanglement swapping [Żuk+93] in a quantum repeater protocol [Bri+98], as shown in Fig. 2.1. Entanglement can also be distributed using quantum network coding protocols [BH20]. The ability to perform teleportation in constant depth immediately gives routing speedups over swap-based methods and even over previous unitary quantum routing methods, since teleportation can be used to quickly exchange distant pairs of qubits.

We show that using measurement and feedback to help prepare long-range entanglement can significantly decrease the time required for routing, even without using a large number of ancillas. In particular, we demonstrate the first superconstant speedup for quantum routing over swap-based routing in the setting where $O(N)$ ancillas are allowed, showing a $O(\log N)$ speedup for the hardest (i.e., worst-case) permutations. Further, our main result proves the first non-trivial limits on the advantage of teleportation-based routing protocols by an $O(\sqrt{N \log N})$ upper bound on the speedup over swap-based routing. Finally, we also show a new swap-based algorithm for sparse routing of k qubits on a graph G in time $O(k + \mathrm{diam}(G))$, where $\mathrm{diam}(G)$ is the diameter

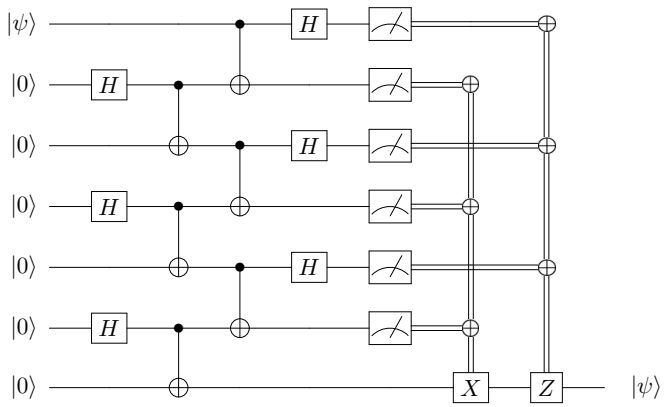


Figure 2.1: Constant-depth long-range teleportation protocol on a path of 7 qubits. The X and Z gates are classically controlled by the parities of the two sets of measurement results. This protocol can be extended to paths of any length without increasing the circuit depth.

of G (i.e, the maximum shortest-path distance between any pair of vertices).

LOCC has previously been useful to give low-depth implementations of specific unitaries, such as quantum fanout [PS13], long-range operations on the surface code [BKS22], and the preparation of a wide range of entangled states [PSC21; Eld+20; Tan+24b; VTV22]. In fact, previous work showed routing speedups by using ancillas [Her20] and by employing LOCC [Ros13]. Using teleportation, Rosenbaum [Ros13] showed a protocol that implements any permutation in constant depth. However, Rosenbaum's protocol uses $O(N^2)$ qubits to perform permutations on $O(N)$ qubits, so that only a negligible fraction of the qubits are data qubits. Engineering qubits is difficult, so it is preferable to use as many of them as possible as data qubits to enable larger computations. Therefore, in this work we consider a more modest $O(1)$ ancillas per data qubit (i.e. there are $O(N)$ ancillas in total). The availability of $O(1)$ ancillas per data qubit is natural in some quantum systems, such as in NV center qubits [Dut+07], quantum dots [LD98], and trapped ions [Bru+19a]. To our knowledge, this is the first work to study quantum routing with

measurement and feedback in the restricted ancilla setting. By studying routing in this regime, we make progress on an open question posed by Herbert [Her20], asking to what extent ancillas can be used to accelerate routing.

Routing is more powerful than state transfer and entanglement distribution [Bos03; Chr+05]. For example, routing qubits from locally prepared Bell states can be used to generate long-range entanglement. The upper bounds in our work therefore also apply to these tasks.

Our work may be of interest to experimental efforts in systems which allow for mid-circuit measurements. In particular, the non-locality enabled by measurement and feedback makes large distances between qubits (i.e, large diameter connectivity graphs) less of a challenge for algorithm implementations. Additionally, knowledge of (teleportation) routing may inform choices of connectivity in systems with these features. In particular, our upper bounds can be used to compare routing overheads on different architectures based on their spectral and isoperimetric properties.

Furthermore, teleportation routing can also provide large advantages for specific permutations, which makes it useful for efficient implementations of algorithms on near-term architectures. This is also of relevance to fault-tolerant quantum computation, as a major obstacle to the implementation of promising quantum error-correcting codes, such as qLDPC codes [BE21; PK21], is their need for long-range syndrome measurements [Hon+24; DBT21]. This can be alleviated by using teleportation to route together distant qubits from each syndrome. Further, protocols to prepare code states and implement logical operations in locality-restricted architectures are constrained by Lieb-Robinson bounds. Recent work [Fri+22] has shown how the use of measurements can accelerate such tasks. Routing schemes enabled by the use of teleportation can also be considered on fault-tolerant architectures, such as, for example, to perform logical

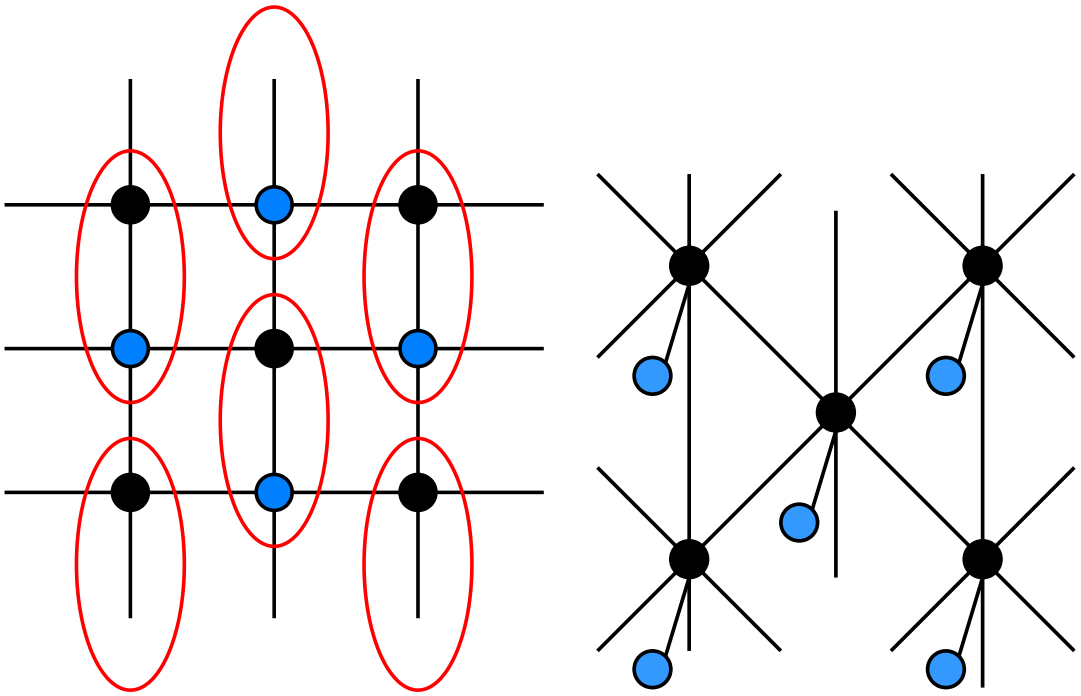
circuits across surface code patches. More generally, the use of measurement and feedback enables speedups from the ability to implement long-range interactions quickly, which can also make algorithms much easier to run on near term architectures [PS13].

2.2 Preliminaries

We consider architectures consisting of N data qubits connected according to a simple graph G (with vertex set $V(G)$ and (undirected) edge set $E(G)$), where an edge $(u, v) \in E(G)$ represents a connection between qubits $u, v \in V(G)$, and $|V(G)| = N$. We consider only *connected* graphs, i.e., graphs in which there is a path from any vertex to any other vertex.

We assume there are a constant number of ancillary qubits per data qubit that can interact only with the data qubit. Further, we assume that disjoint two-qubit gates can be performed between adjacent qubits in depth 1. Up to a constant overhead, this is equivalent to having fast (instantaneous) ancilla interactions since any unitary on the data qubit and ancillas can be decomposed into a constant number of two-qubit gates. As our results are asymptotic, they are insensitive to a constant overhead.

Ancillary qubits corresponding to different data qubits are not directly connected. However, gates between ancillary qubits of neighboring vertices can be performed in depth 1 by swapping ancillas with their corresponding data qubits, performing the desired 2-qubit gate between data qubits, and swapping again with the ancillas. This model can be implemented in realistic quantum architectures with attached ancillas [LD98; Dut+07; Bru+19a] as well as architectures with grid connectivity such as superconducting qubits [Nat+21; Aru+19]. For example, Fig. 2.2a shows an architecture where ancillas are interspersed with data qubits on a grid. This can be represented



(a) (b)

Figure 2.2: (a) A grid architecture with ancillas (blue) interspersed between data qubits (black). The red ovals indicate which ancilla corresponds to each data qubit. (b) An equivalent architecture in our model.

in our model as Fig. 2.2b. Both models are equivalent and can simulate each other with only constant depth overhead.

The task of routing involves permuting data qubits on the graph. We use the notation

$$\{(1, \pi(1)), (2, \pi(2)), \ldots, (N, \pi(N))\} \tag{2.1}$$

to denote a permutation on N vertices, where $\pi(i)$ is the vertex to which we must move the ith qubit. We also write

$$[N] \coloneqq \{1, 2, \ldots N\}. \tag{2.2}$$

We consider the following models of routing.

1. Swap routing: In this model, the only allowed gates between adjacent qubits are swap gates.

2. LOCC routing: In this model, we are allowed to perform arbitrary 2-qubit gates on disjoint pairs of qubits in a single time step. Further, in the same time step, we are allowed to perform single-qubit measurements (on data and ancilla qubits) and adaptively apply arbitrary single-qubit gates. We refer to this as *fast measurement and feedback*. Gates in later time steps can be applied adaptively, conditioned on all previous measurement results.

3. Teleportation routing: In this model, data qubits can be teleported along disjoint paths to ancilla registers at arbitrary distances in depth 1. Using this ability, a swap between the ends of a path can be performed in constant depth. Note that the qubits along a teleportation path cannot be involved in any other operations during a round of teleportation. However, teleportation between multiple pairs of qubits can be performed in parallel if there exist paths for each pair that have no more than a constant number of intersections per vertex, since we allow a constant number of ancilla qubits per data qubit. This model is a specialization of LOCC routing as the ability to perform fast measurement and feedback allows us to perform quantum teleportation, transporting a single qubit to any vertex in constant depth. The entanglement required for quantum teleportation is produced using an entanglement swapping protocol [Żuk+93], as depicted in Fig. 2.1. A swap between the ends of a path can be performed by teleporting the qubit at each end to the opposite end, or by performing gate teleportation [GC99] of a swap gate.

We are particularly interested in the *routing time* $\mathrm{rt}(G, \pi)$, which is the minimum circuit depth to perform the permutation π on the data qubits of G. The worst-case routing time of a

graph G is

$$\mathrm{rt}(G) \coloneqq \max_{\pi \in \mathcal{S}_N} \mathrm{rt}(G, \pi) \tag{2.3}$$

where $\mathcal{S}_N$ is the symmetric group, i.e., the group of all permutations of N elements. We let $\mathrm{rt}_{\mathrm{tele}}(G)$ denote the routing time in the teleportation model, $\mathrm{rt}_{\mathrm{LOCC}}(G)$ denote the routing time in the LOCC model, and $\mathrm{rt}_{\mathrm{swap}}(G)$ denote the routing time in the swap model.

2.3 Bounds on routing time

In this section, we discuss known bounds on the routing time for both swap and LOCC routing.

2.3.1 Lower bounds

If a permutation can only be implemented by sending a large number of tokens through a small number of vertices, then any circuit for performing it must have high depth, since each vertex can only hold one token at a time. This gives a natural lower bound on the routing time. To formalize this, we consider the *vertex expansion* (or *vertex isoperimetric number*) $c(G)$ of a graph G, defined as follows.

Definition 2.1. The vertex expansion of a graph G is

$$c(G) \coloneqq \min_{X \subseteq V(G)} \frac{|\delta X|}{\min\{|X|, |\overline{X}|\}}, \tag{2.4}$$

where

$$\overline{X} = V(G) - X \tag{2.5}$$

is the complement of X, and

$$\delta X = \{v \in \overline{X} \mid \exists\, u \in X \text{ s.t. } (u, v) \in E(G)\} \tag{2.6}$$

is the vertex boundary of X.

Note that $c(G) \leq 1$:

$$\begin{aligned}
&\min_{X \subseteq V(G)} \frac{|\delta X|}{\min\{|X|, |\overline{X}|\}} && (2.7)\\
&= \min_{X \subseteq V(G)} \left(\frac{|\delta X|}{\min\{|X|, |\overline{X}|\}}, \frac{|\delta \overline{X}|}{\min\{|X|, |\overline{X}|\}} \right) && (2.8)\\
&= \min_{X \subseteq V(G)} \frac{\min(|\delta X|, |\delta \overline{X}|)}{\min\{|X|, |\overline{X}|\}} && (2.9)\\
&\leq 1. && (2.10)
\end{aligned}$$

In addition, for a connected graph, since $|\delta X| \geq 1$ and $\min\{|X|, |\overline{X}|\} \leq \frac{N}{2}$ for any X, we have $c(G) \geq \frac{2}{N}$ Therefore, for connected graphs, $c(G) \in \left[\frac{2}{N}, 1\right]$.

Any connected simple graph G satisfies the following.

Theorem 2.2 (Isoperimetric lower bound [Bap+23])**.**

$$\mathrm{rt}_{\mathrm{LOCC}}(G) \geq \frac{2}{c(G)} - 1. \tag{2.11}$$

Since $\mathrm{rt}_{\mathrm{swap}}(G) \geq \mathrm{rt}_{\mathrm{tele}}(G) \geq \mathrm{rt}_{\mathrm{LOCC}}(G)$, this lower bound applies to swap- and teleportation-based routing as well.

We can also lower bound the swap-based routing time by the diameter of the graph (i.e,

the maximum shortest-path distance between any pair of vertices) since swapping two vertices at distance d requires a swap circuit of depth at least d.

Theorem 2.3 (Diameter lower bound)**.**

$$\mathrm{rt}_{\mathrm{swap}}(G) \geq \mathrm{diam}(G). \tag{2.12}$$

Note that this bound does *not* apply to teleportation or LOCC routing.

2.3.2 Upper bounds

On any graph, a classical swap algorithm can route on an N-vertex tree in depth $O(N)$ [Zha99]. Recall that we only consider connected graphs, so we can always route on a spanning tree with swaps in depth $O(N)$. We thus have the following upper bounds.

Theorem 2.4. *For any N-vertex connected graph G,*

$$\mathrm{rt}_{\mathrm{swap}}(G) = O(N) \tag{2.13}$$

This bound also implies that $\mathrm{rt}_{\mathrm{tele}}(G) = O(N)$ and $\mathrm{rt}_{\mathrm{LOCC}}(G) = O(N)$.

We can prove a tighter bound for *sparse* routing. Let $\mathrm{rt}_{\mathrm{swap}}(G, k)$ denote the worst-case routing time on G over permutations that move at most k tokens. Using reversals, [Bap+23] gives a routing algorithm that takes depth $O(\mathrm{diam}(G) + k^2)$. We improve this result, using swaps with ancillas, to show the following.

Theorem 2.5 (Sparse routing)**.** *For any N-vertex connected simple graph G and $k \in [N]$,*

$$\mathrm{rt}_{\mathrm{swap}}(G, k) = O(\mathrm{diam}(G) + k). \tag{2.14}$$

Proof sketch. Call all tokens v with $\pi(v) \neq v$ *marked*. There are k marked tokens. There are three main steps in our algorithm:

1. Hide all unmarked tokens in the ancillas by performing swaps. Route the k marked tokens to span a tree subgraph in time $O(\mathrm{diam}(G))$.

2. Permute the k tokens on the tree subgraph, using the procedure from [Zha99], in time $O(k)$.

3. Reverse the first step, thereby moving the k tokens from the subgraph to the appropriate target locations in time $O(\mathrm{diam}(G))$. Restore the unmarked tokens from the ancillas.

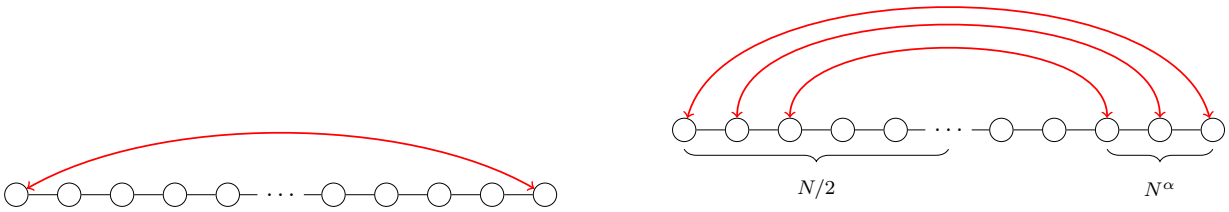
See Appendix A.1 for the full proof. □

2.4 Faster permutations with teleportation

The ability to perform teleportation immediately suggests possibilities for speedups over swap-based routing. Swap-based routing must obey the diameter lower bound (Theorem 2.3), so permutations that involve long-range swaps (e.g., between diametrically separated pairs of vertices) should be sped up by teleportation.

We define the *teleportation advantage* for a specific permutation to quantify this speedup:

$$\mathrm{adv}(G, \pi) := \frac{\mathrm{rt}_{\mathrm{swap}}(G, \pi)}{\mathrm{rt}_{\mathrm{tele}}(G, \pi)}. \tag{2.15}$$



(a) Diameter-length permutation π_{diam} (shown by the red double-sided arrow) on the path graph P_N (shown in black).

(b) Rainbow permutation $\pi^{\alpha}_{\text{rainbow}}$ (shown by red double-sided arrows) on the path graph P_N (shown in black).

Figure 2.3: Permutations on a 1D lattice

We now consider the following permutation on the path graph P_N: $\pi_{\text{diam}} = \{(1, N), (2, 2), \ldots, (N-1, N-1), (N, 1)\}$ (see Fig. 2.3a).

By the diameter lower bound, this permutation takes depth $\Omega(N)$ with swaps. However, with teleportation it takes depth $O(1)$, showing that $\text{adv}(P_N, \pi_{\text{diam}}) = \Omega(N)$.

This further generalizes to permutations that require multiple long-range swaps. For example, consider a *rainbow* permutation $\pi^{\alpha}_{\text{rainbow}}$, as depicted in Fig. 2.3b. This permutation involves performing N^{α} swaps across a 1D lattice for some $\alpha \in [0, 1]$. With swaps, this takes depth $\Theta(N)$ by the diameter bound, but with teleportation it takes depth N^{α}, by a procedure that simply teleports each pair into place sequentially. This gives a polynomial advantage: $\text{adv}(P_N, \pi^{\alpha}_{\text{rainbow}}) = O(N^{1-\alpha})$.

These permutations allow speedups bounded by the diameter of the graph. Any single teleportation step can be simulated by swaps in depth $O(\text{diam}(G))$, by simply swapping along the shortest path between the initial qubit and the final destination. Intuitively, one might therefore expect that teleportation routing could achieve at most a diameter-factor speedup. However, there exist some graphs and permutations for which we can obtain even larger speedups. Teleportation speedups are not limited by the graph diameter since teleportation protocols can utilize multiple

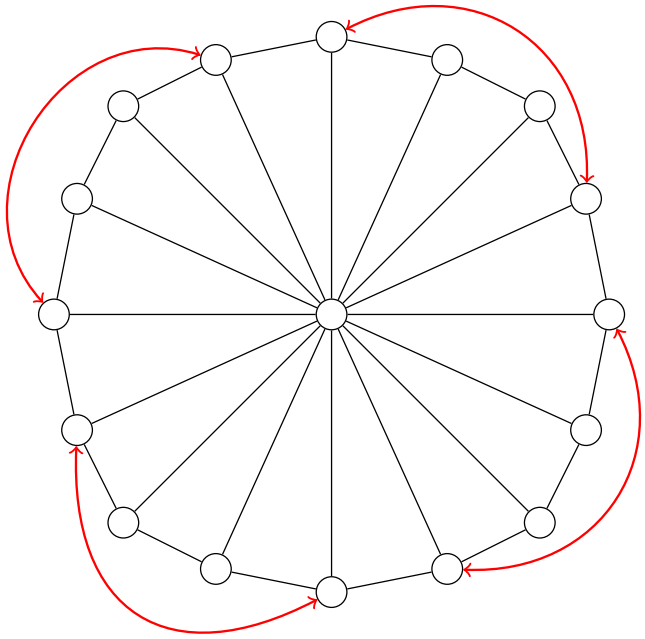


Figure 2.4: Permutation π^l_{wheel} (shown by red double-sided arrows) that exchanges l pairs of vertices on the wheel graph W_{N+1} (shown in black).

longer paths together to avoid intersections.

To illustrate this, consider the example of a *wheel* graph W_{N+1}, as shown in Fig. 2.4. The $(N+1)$-vertex wheel graph, with central vertex $N+1$, has edges

$$E(W_{N+1}) = \{(u,v) \mid u - v = 1 \pmod N \text{ or } v = N+1\}. \tag{2.16}$$

The diameter of W_{N+1} is 2. On this graph, consider the permutation (shown in red in Fig. 2.4)

$$\begin{aligned}\pi^l_{\text{wheel}} \coloneqq \{&(1, N/l), (N/l+1, 2N/l), \ldots, \\ &(N - N/l + 1, N)\}\end{aligned} \tag{2.17}$$

that exchanges l pairs of vertices spaced along the "rim" of the wheel (assume $l \mid N$). For swap-

based algorithms, this can be done in depth $\min\{3l, N/l - 1\}$ by routing the qubits sequentially through the central vertex or routing them in parallel along the "rim", whichever is faster.

This is optimal up to constant factors, by the following reasoning. If there exists a data token that does not pass through the central node, the routing time must be at least $N/l - 1$, which is the travel distance along the rim. On the other hand, if every data token passes through the central node, then there must be at least $2l$ steps in the algorithm. Therefore

$$\mathrm{rt}_{\mathrm{swap}}(W_{N+1}, \pi^{l}_{\mathrm{wheel}}) \geq \min\{2l, N/l - 1\}. \tag{2.18}$$

However, in the teleportation routing model, this permutation can be performed in constant depth by performing l teleportations in parallel along non-intersecting paths on the wheel rim. Therefore,

$$\mathrm{rt}_{\mathrm{tele}}(W_{N+1}, \pi^{l}_{\mathrm{wheel}}) = O(1). \tag{2.19}$$

Setting $l = \sqrt{N/2}$, we obtain a maximum teleportation advantage $\mathrm{adv}(W_{N+1}, \pi^{l}_{\mathrm{wheel}}) = \Theta(\sqrt{N})$ for this class of permutations, even though $\mathrm{diam}(W_{N+1}) = O(1)$. Teleportation therefore enables super-diametric speedups.

2.5 Teleportation Advantage

While π_{diam}, $\pi^{\alpha}_{\mathrm{rainbow}}$, and π^{l}_{wheel} allow for teleportation speedups, they are not the worst-case permutations on their respective graphs. For example, consider the full reflection on the line graph, i.e., a rainbow permutation with $\alpha = 1$. This permutation requires depth $\Theta(N)$ for both swap- and teleportation-based routing. Similarly, on the wheel graph with an even number

of vertices, the permutation π with $\pi(i) = i + \lfloor N/2 \rfloor \bmod N$ for all $i \in [N]$ requires depth $\Theta(N)$ for both types of routing as well. Thus, although these graphs have teleportation speedups for specific permutations, there is no separation between their swap and teleportation routing numbers.

To compare the relative strength of the teleportation routing model to the swap-based routing model for all permutations, we aim to understand how much teleportation improves worst-case permutations. We measure the relative strength of the teleportation model by the separation in teleportation and swap-based routing numbers, which we define as the *worst-case teleportation advantage*:

$$\mathrm{adv}(G) \coloneqq \frac{\mathrm{rt}_{\mathrm{swap}}(G)}{\mathrm{rt}_{\mathrm{tele}}(G)}. \tag{2.20}$$

Note that this is *not* the worst-case ratio of routing numbers for a single specific permutation, i.e., $\mathrm{adv}(G)$ is not necessarily the same as $\max_\pi \mathrm{adv}(G, \pi)$. (Indeed, as discussed above, these two quantities differ for the path and wheel graphs.) Instead, $\mathrm{adv}(G)$ can be thought of as the speedup teleportation provides for the general task of routing on a particular graph in the worst case, rather than for implementing a specific permutation. It also allows us to compare different graphs: teleportation routing offers greater worst-case guaranteed speedups on graphs with higher $\mathrm{adv}(G)$.

It is not immediately obvious that we should expect adv to be greater than 1 for any graph. However, we now describe a graph that does offer a worst-case speedup for teleportation. This graph, which we denote by $L(n)$ (with $N = 2^n - 1$ vertices), has $\mathrm{adv}(L(n)) = n = \log_2(N+1)$.

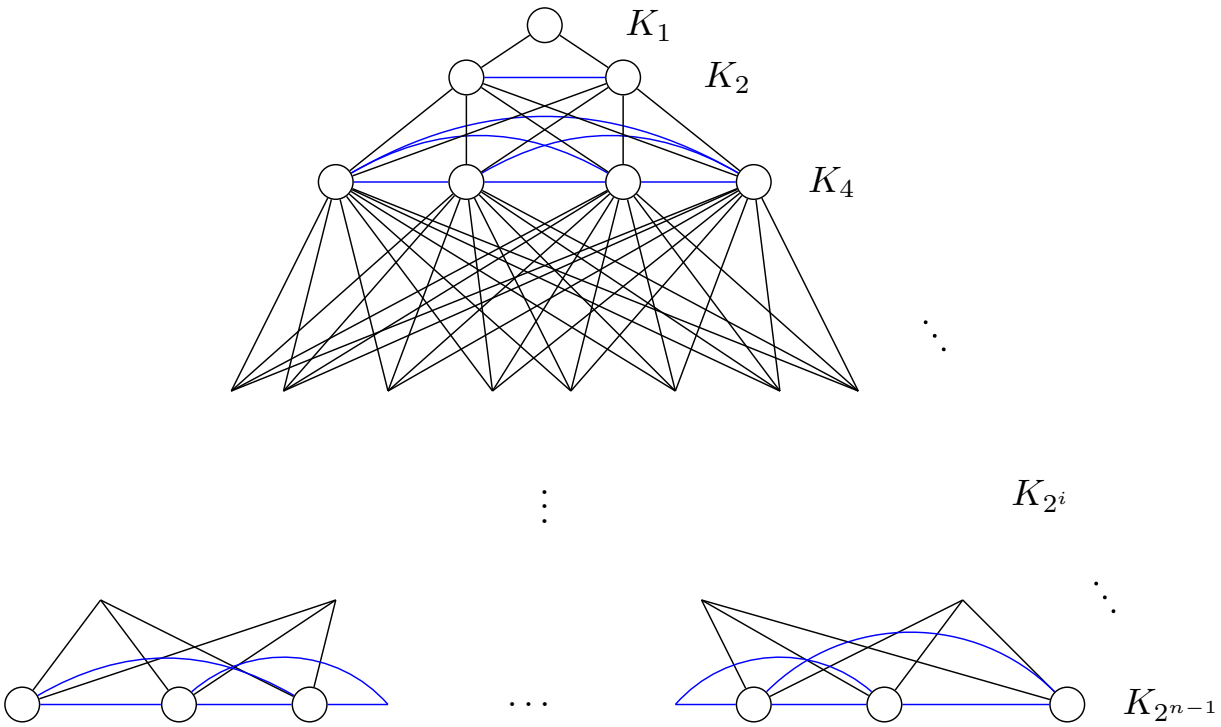


Figure 2.5: The graph $L(n)$. The black lines show edges between layers, while blue lines show edges within a layer (colored for visibility).

The graph $L(n)$ (depicted in Fig. 2.5) has

$$V(L(n)) = \{(r, i) \mid r \in [n], i \in [2^{r-1}]\} \tag{2.21}$$

and

$$\begin{aligned} E(L(n)) = \{((r, i_1), (r, i_2)) \mid & \\ r \in [n],\ i_1 < i_2 \in [2^{r-1}]\} & \\ \cup \{((r_1, i_1), (r_2, i_2)) \mid & \\ r_2 - r_1 = 1,\ i_1, i_2 \in [2^{r-1}]\}. & \end{aligned} \tag{2.22}$$

In words, $L(n)$ is a ladder formed by arranging complete graphs K_{2^k} for $k \in \{0, 1, \ldots, n-1\}$ in

horizontal layers, and then connecting every vertex in a given layer with every vertex one layer above or below. The total number of vertices in this graph is

$$N = \sum_{k=0}^{n-1} 2^k = 2^n - 1. \tag{2.23}$$

The diameter of $L(n)$ is exactly $n - 1 = \log_2(N + 1) - 1$. Theorem 2.3 then implies

$$\mathrm{rt}_{\mathrm{swap}}(L(n)) = \Omega(\log(N)). \tag{2.24}$$

With teleportation, we show that routing can be performed in depth $O(1)$. The key idea behind the teleportation protocol is that every layer of $L(n)$ has one more node than all the layers above it together. This allows us to identify a unique node in each layer corresponding to any node from a higher layer. We can then route tokens by simply teleporting along the path formed by the unique nodes from each layer, corresponding to the source vertex of the token to be routed.

This teleportation routing procedure establishes the following.

Proposition 2.6. $\mathrm{rt}_{\mathrm{tele}}(L(n)) = O(1)$.

Proof. For any permutation $\pi \in S_{2^n-1}$, we construct a set of paths $\{P(u, \pi(u)) \mid u \in V(L(n))\}$ between every node and its destination such that each vertex of the graph belongs to at most four paths in the set.

Label every vertex in the graph with an n-bit address as follows. To every node in the subgraph K_{2^i} (corresponding to layer $i + 1$ of the ladder), assign a unique integer u in the range $[2^i, 2^{i+1} - 1]$. (Since the layer is a complete graph, the order within a layer is arbitrary.) Equivalently, we may refer to node u by its binary representation $b(u)$, which is an $(i + 1)$-bit string

with a leading 1, i.e., of the form $b(u) = (1\ldots)$.

For any vertex $u \in V$, define $r(u, i)$ to be the vertex whose address is $(10^{i-1}b(u))$, i.e., the address of u appended to a leading 1 i places to the left. Note that $r(u, i+1)$ is adjacent to $r(u, i)$ and lies in the layer immediately below $r(u, i)$. Define $r(u, 0) = u$.

Now, given two vertices u, v separated by a distance d, define a *canonical path* $P(u, v) = P(v, u)$ as the sequence of the following nodes: $(u, r(u, 1), \ldots, r(u, d-1), v)$, where we assume $u < v$ without loss of generality. If $d = 1$, then $P(u, v) = (u, v)$. We now show that for any permutation $\pi \in S_{2^n-1}$, the set of canonical paths $\{P(u, \pi(u)) \mid u \in V(L(n))\}$ intersects any vertex at most four times.

Fix an arbitrary vertex v. By construction, v lies in $P(v, \pi(v))$ and $P(v, \pi^{-1}(v))$. Now suppose a path $P(u, \pi(u))$ passes through $v \notin \{u, \pi(u)\}$. Then either $u < v < \pi(u)$ or $\pi(u) < v < u$. Without loss of generality, we assume the former. Since $P(u, \pi(u))$ is canonical, $b(v) = (10^i b(u))$ for some $i \geq 0$. Suppose a different path $P(u', \pi(u'))$ also intersects v. Then there are two cases to consider: $u' < \pi(u')$ and $u' > \pi(u')$. In the first case, $b(v) = (10^{i'} b(u'))$. This is only possible when $i = i'$ and $u = u'$, which implies that $P(u', \pi(u')) = P(u, \pi(u))$ (giving one intersecting path at v). In the second case, the same reasoning implies that $u = \pi(u')$. In this case, there are two intersecting paths $P(u, \pi(u))$ and $P(\pi^{-1}(u), u)$ at v. Therefore, in addition to $P(v, \pi(v))$ and $P(v, \pi^{-1}(v))$, at most two other paths can intersect at v, giving a total of at most four paths.

Finally, construct one Bell pair for every edge in every canonical path $P(u, \pi(u))$, using distinct local ancillas for every pair. The number of Bell pairs shared at any vertex is at most $6 = O(1)$, requiring 6 local ancillas per vertex. Using the standard repeater protocol (Fig. 2.1) along each canonical path, one can then carry out simultaneous teleportation of all data qubits to

their destination vertices $v \mapsto \pi(v)$ in constant depth. Therefore, any permutation of the qubits can be implemented in depth $O(1)$. □

2.6 Bounding the Teleportation Advantage

In the previous section, we described a graph with logarithmic teleportation advantage. In this section, we examine limits on the teleportation advantage. In order to understand the power of teleportation in general, we specifically aim to bound the *maximum teleportation advantage*

$$\mathrm{adv}^* \coloneqq \max_G \mathrm{adv}(G) \tag{2.25}$$

over all graphs with a fixed number of vertices. This quantity measures the maximum speedup teleportation can provide on worst-case permutations for *any* graph. We also show tighter bounds on the advantage for some common classes of graphs.

We immediately have an upper bound on adv from Theorem 2.4. Since any teleportation algorithm must have depth $\Omega(1)$, and a swap algorithm can implement any permutation in depth $O(N)$, we have

$$\mathrm{adv}^* = O(N). \tag{2.26}$$

We now show a tighter bound.

2.6.1 Advantage for general graphs

Combining Theorem 2.2 and Theorem 2.3, we have

$$\mathrm{rt}_{\mathrm{swap}}(G) \geq \max\{\frac{2}{c(G)} - 1, \mathrm{diam}(G)\} \tag{2.27}$$

We now consider the relationship between $\mathrm{diam}(G)$ and $\frac{1}{c(G)}$. Intuitively, increasing the diameter while keeping N constant 'stretches' the graph, tightening bottlenecks. This causes $c(G)$ to decrease. Similarly, eliminating bottlenecks in the graph requires adding more edges across cuts, thereby increasing the connectivity of the graph and reducing the diameter. We thus expect that graphs with higher diameter will have higher $\frac{1}{c(G)}$, and graphs with small $\frac{1}{c(G)}$ will have small diameter. We can express this relation more precisely as follows.

Lemma 2.7. *For any connected simple graph G,*

$$\mathrm{diam}(G) \leq 2\frac{\log \frac{N}{2}}{\log\left(1 + c(G)\right)} + 2. \tag{2.28}$$

Proof. See Appendix A.2. □

One might expect graphs with large diameter to allow large speedups, since the diameter lower bound only applies to swap routing. However, as illustrated by Lemma 2.7, graphs with large diameter also have tight bottlenecks, and therefore, by Theorem 2.2, are not likely to permit large speedups.

We now show our main results bounding the advantage. Our main technical result bounds the advantage in terms of the diameter of the graph. We note that this bound also applies to the

separation between swaps and teleportation routing for any permutation, and not just the worst-case separation.

Lemma 2.8. $\mathrm{adv}(G) = O(\sqrt{N} + \mathrm{diam}(G))$.

Proof. We construct a swap-based protocol that can simulate a single round of teleportation in depth $O(\sqrt{N} + \mathrm{diam}(G))$, thereby upper bounding the teleportation advantage.

A single round of a teleportation protocol performs teleportation along a set of paths. These paths must intersect no more than a constant number of times per vertex, since there are only a constant number of ancillas per vertex.

For all paths from the teleportation protocol of length at most $\sqrt{N}$, we swap along the paths in parallel. Since each vertex only has a constant number of paths going through it, a qubit can move through every vertex in constant depth. Therefore, these swaps can be performed in depth $O(\sqrt{N})$.

For an N-vertex graph, the number of paths of length at least l that intersect at most a constant number of times is in $O(N/l)$. Therefore, since each long path corresponds to a single token, after routing along all paths of length at most $\sqrt{N}$, we have $O(\sqrt{N})$ tokens left to route. By Theorem 2.5, this can be done in depth $O(\sqrt{N} + \mathrm{diam}(G))$.

We can thus simulate each teleportation round in depth $O(\sqrt{N} + \mathrm{diam}(G))$, which completes the proof. □

Combining our results, we now have a bound on the advantage for *any* graph.

Theorem 2.9. $\mathrm{adv}^* = O(\sqrt{N \log N})$.

Proof. First, combining Theorem 2.4 and Theorem 2.2, we have

$$\mathrm{adv}(G) = O(N \cdot c(G)). \tag{2.29}$$

Combining this bound with the bound from Lemma 2.8, we have

$$\mathrm{adv}(G) \leq \min\left\{O(N \cdot c(G)), O(\sqrt{N} + \mathrm{diam}(G))\right\}. \tag{2.30}$$

We know that $c(G) > 0$. Using the fact that $\log(x) \geq 1 - 1/x$ for $x > 0$, we have

$$\frac{1}{\log(1 + c(G))} \leq \frac{1}{c(G)} + 1 = O\left(\frac{1}{c(G)}\right), \tag{2.31}$$

where in the last equality we used $c(G) \leq 1$. Applying this to Lemma 2.7 and Eq. (2.30), we have

$$\mathrm{adv}(G) \leq \min\left\{O(N \cdot c(G)), O\left(\sqrt{N} + \frac{\log(N)}{c(G)}\right)\right\}. \tag{2.32}$$

Recall the definition of the maximum teleportation advantage from Eq. (2.25):

$$\mathrm{adv}^* := \max_G \mathrm{adv}(G). \tag{2.33}$$

Therefore,

$$\begin{aligned}\mathrm{adv}^* \leq \max_G \min\Big\{&O(N \cdot c(G)),\\ &O\left(\sqrt{N} + \frac{\log(N)}{c(G)}\right)\Big\}.\end{aligned} \tag{2.34}$$

As $c(G)$ varies, the two bounds in the minimum vary inversely. The first bound, from Eq. (2.29),

is monotonically increasing in $c(G)$ for $c(G) \in (0, 1]$. The second bound is monotonically decreasing in $c(G)$ for $c(G) \in \left(0, \frac{\log N}{\sqrt{N}}\right]$. Note that when $c(G) \sim 1/N$ (recall that $c(G) \geq 2/N$), the first bound is smaller, while when $c(G) \sim \frac{\log N}{\sqrt{N}}$, the second bound is smaller. The largest minimum of the two bounds is thus obtained when they are equal.

The minimum of the two bounds is thus maximized when $c(G) = \sqrt{(\log N)/N}$. Note that even if a graph with $c(G) = \sqrt{(\log N)/N}$ does not exist, any other value of $c(G)$ will result in a smaller right-hand side of Eq. (2.32). With $c(G) = \sqrt{(\log N)/N}$, we obtain

$$\mathrm{adv}^* = O\left(\sqrt{N \log N}\right) \tag{2.35}$$

as claimed. □

This bound applies to any graph, and is thus independent of the diameter of the graph. Therefore, this result shows that in graphs with diameter $\omega(\sqrt{N \log N})$, we cannot obtain a routing time separation between teleportation- and swap-based routing that is proportional to the diameter.

Next we show tighter bounds for a few common families of graphs.

2.6.2 Grids

For d-dimensional grids (i.e, $P_n^{\square d}$, the d-fold Cartesian product of the path graph P_n, with $N = n^d$ vertices), the vertex cut bound (Theorem 2.2) gives

$$\mathrm{rt}_{\mathrm{LOCC}}(P_n^{\square d}) \geq \frac{2}{c(P_n^{\square d})} - 1 \geq n - 1, \tag{2.36}$$

where $c(P_n^{\square d}) \leq 2/n$ follows from considering a hyperplane that bisects the grid along one dimension. From [ACG94], we have

$$\mathrm{rt}_{\mathrm{swap}}(G_1 \square G_2) = 2\,\mathrm{rt}_{\mathrm{swap}}(G_1) + \mathrm{rt}_{\mathrm{swap}}(G_2). \tag{2.37}$$

Therefore, the swap routing time of a d-dimensional grid is $O(dN^{1/d}) = O(dn)$. For constant d, this saturates the cut bound in Eq. (2.36). Therefore, there is no worst-case speedup from either teleportation or full LOCC, i.e, $\mathrm{adv}(P_n^{\square d}) = 1$.

2.6.3 Expander graphs

We bound the advantage for spectral expander graphs to be $\mathrm{poly}(\log N)$. The *(normalized) Laplacian* of a graph, G, is defined as

$$\mathcal{L}_{u,v} = \begin{cases} 1 & \text{if } u = v \\ -\frac{1}{\sqrt{d_v d_u}} & \text{if } (u,v) \in E(G) \\ 0 & \text{otherwise,} \end{cases} \tag{2.38}$$

where d_v is the degree of vertex v. The matrix $\mathcal{L}$ is symmetric and positive such that we can order its eigenvalues as $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$. We write $\lambda(G)$ for λ_1 of the Laplacian of G. *Spectral expander graphs* are graphs of bounded degree with $\lambda(G) = \Omega(1)$. For a comprehensive introduction to spectral graph theory, consult [Chu96].

To bound the advantage for spectral expander graphs, we first use the following upper bound on the swap-based routing number. Let $d_* := \frac{\max_{v \in V} d_v}{\min_{v \in V} d_v}$ denote the degree ratio of a graph.

Theorem 2.10 ([Bap+23]). *For any graph G and permutation π,*

$$\mathrm{rt}_{\mathrm{swap}}(G,\pi) = O\left(\frac{d_*}{\lambda(G)^2}\log^2 N\right). \tag{2.39}$$

Combining this result with the lower bound of Theorem 2.2, we immediately get

$$\mathrm{adv}(G) = O\left(\frac{d_* c(G)\log^2 N}{\lambda(G)^2}\right). \tag{2.40}$$

Thus graphs with $\lambda(G) = \Omega(1)$ and $d_* = O(1)$ (such as spectral expanders) have at most a polylogarithmic advantage.

2.6.4 Hypercubes

The swap-based routing time for a d-dimensional hypercube Q_d is [ACG94; LLY10]

$$\mathrm{rt}_{\mathrm{swap}}(Q_d) = \Theta(d). \tag{2.41}$$

Since $|V(Q_d)| = N = 2^d$,

$$\mathrm{rt}_{\mathrm{swap}}(Q_d) = \log N. \tag{2.42}$$

Now, we will show that $c(Q_d) = \Theta(\frac{1}{\sqrt{d}})$. In a hypercube, Hamming balls (i.e., sets of all points with Hamming weight $\leq r$ for some integer r) have the smallest boundary of all sets of a given size [Har66]. Taking the Hamming ball of radius $d/2$ as X, we have $|X| = 2^{d-1}$ and $|\delta X| = \binom{d}{d/2} = \Theta(2^d/\sqrt{d})$. Therefore, $c(G) = \Theta(1/\sqrt{d})$. Using Theorem 2.2, we have $\mathrm{rt}_{\mathrm{tele}}(G) = \Omega(\sqrt{d}) = \Omega(\sqrt{\log N})$. Teleportation thus offers at most an $O(\sqrt{\log N})$ advantage on hypercubes.

2.6.5 Other graphs

The cyclic butterfly graph B_r has been proposed as a constant-degree interaction graph that allows for fast circuit synthesis [Cow+19; Bri17]. Each of the $N = r2^r$ vertices is labelled $(w, i) \in \{0,1\}^r \times [r]$. Vertices (w, i) and $(v, i+1 \bmod r)$ are connected if $w = v$ or if w and v differ by exactly one bit in the ith position. The cyclic butterfly has diameter $O(\log N)$, degree 4, and $\mathrm{rt}_{\mathrm{swap}}(B_r) = O(\log N)$ [Bri17].

We now show that the $O(\log N)$ protocol is optimal even for teleportation routing on the cyclic butterfly graph, so $\mathrm{adv}(G) = O(1)$. Bipartition the vertices into sets $X, \overline{X}$ such that X consists of all rows with bit $j = 0$ for some j, and $\overline{X}$ consists of all rows with bit $j = 1$. For this partition, $|X| = r2^{r-1}$ and $|\delta X| = 2^r$, so $c(G) \leq 2/r$. Since $r = \Theta(\log N)$, $c(G) = O(\frac{1}{\log N})$, so from Theorem 2.2, $\mathrm{adv}(G) = O(1)$.

The complete graph K_N has $\mathrm{rt}_{\mathrm{swap}}(K_N) = O(1)$, and therefore has $\mathrm{adv}(K_N) = 1$.

Finally, graphs with poor expansion properties—in particular, with vertex expansion $c(G) = O(\frac{\mathrm{poly}(\log N)}{N})$—have at most polylogarithmic advantage by Eq. (2.29).

2.7 Discussion

In this work, we have used quantum teleportation to speed up the task of permuting qubits on graphs. We have shown examples of specific types of permutations that can be sped up by teleportation. Further, we have shown an example of a graph that exhibits a worst-case teleportation routing speedup of $\log N$. Our main technical result (Theorem 2.9) is a general upper bound of $O(\sqrt{N \log N})$ on the worst-case routing speedup. We also show that many practical architectures cannot implement arbitrary interactions with low overhead, even with fast LOCC (unlike

previous work which only considered unitary evolution). Such a negative result provides useful constraints for the design of quantum devices, suggesting that designing new architectures may prove fruitful.

Our work leaves an open question on whether there exists a graph with $\mathrm{adv}(G) = \omega(\log N)$. Such a graph cannot be a spectral expander graph as per Theorem 2.10. From Lemma 2.7, we know that a graph with large diameter will have poor expansion properties (small $c(G)$) and therefore will not have a large teleportation advantage as per Eq. (2.29). Some candidate graphs for a superlogarithmic teleportation advantage are those with $c(G) \approx \sqrt{(\log N)/N}$. Such graphs may come closer to achieving a teleportation advantage given by the upper bound of Theorem 2.9.

Furthermore, we believe that there should exist a tighter upper bound than Theorem 2.9 on the maximum teleportation advantage for any graph. This is one particularly interesting direction in resolving the advantage of a teleportation protocol over swaps. There could be more sophisticated methods that give tighter bounds by exploiting parallelism. A possible approach to tightening this bound would be to show a swap protocol that performs routing from multiple teleportation rounds in parallel, since swap paths need not obey the strict conditions of teleportation paths (namely, allowing only a constant number of path intersections per vertex).

We have primarily focused on the teleportation model of routing. However, teleportation routing is a special case of the more general LOCC model of routing. We currently do not know whether the full power of LOCC can provide a super-constant speedup over teleportation routing. This is analogous to another open question, namely whether routing with arbitrary 2-qubit gates—or even with arbitrary bounded 2-qubit Hamiltonians—can provide a super-constant speedup over swap-based routing [Bap+23].

Herbert [Her20] posed the question of establishing to what extent ancillas can be used to reduce the routing depth. Rosenbaum [Ros13] showed an $O(1)$ routing protocol on N qubits with $O(N^2)$ ancillas (i.e., an advantage of $O(N)$), while systems without ancillas cannot perform LOCC or teleportation routing, and therefore cannot exhibit any speedups. We have investigated an intermediate regime, and have shown that a linear number of ancillas cannot allow for speedups greater than $O(\sqrt{N \log N})$. It remains an open question to further investigate the space-time tradeoff between the number of ancilla qubits and the routing time.

We assume noiseless circuits, but in the presence of noise the performance of teleportation protocols depends directly on the fidelity of the required resource Bell pairs. We are primarily interested in ways to use teleportation for routing, and Bell pairs are necessary for this process. Our current teleportation routing model does not distinguish between routing over long or short paths, but a more comprehensive model of routing could prioritize shorter paths as they will be less error prone without error correction. Alternatively, we could use a purification protocol [Ben+96b] to prepare high-fidelity Bell pairs at the cost of additional ancillas and overhead, or we could encode our state in an error-correcting code [Den+02] to suppress the error rate when operating between nodes. If operating in a quantum network, we can make use of protocols generalizing entanglement swapping from Bell basis measurements to n-qubit GHZ states to improve the performance of repeater protocols in lossy quantum networks [Pat+22]. Alternatively, we can prepare a high-fidelity Bell pair by performing multiple repeater protocols along different paths in parallel [Pan+19] or using multiplexers on each edge [Lee+22].

A more general task than routing is to perform unitary synthesis, i.e, decompose a particular unitary into 2-qubit gates that can be applied on our locality-constrained qubits. It remains an open question to understand how much unitary synthesis can be sped up by using LOCC with a

linear number of ancillary qubits. Previous work has shown an $\Omega(N)$ speedup for implementing fanout [PS13] and preparing GHZ and W states [PSC21], and an $\Omega(\sqrt{N})$ speedup for preparing toric code states [PSC21], which takes time $\Omega(\sqrt{N})$ without LOCC [BHV06]. Previous work has also shown how measurements of cluster states can be used to efficiently prepare long-range entanglement [Tan+24b] and states with exotic topological order [VTV22]. In principle, LOCC could provide superlinear speedups for unitary synthesis, as we currently have no upper bounds on the advantage for arbitrary unitaries.

Chapter 3: Quantum routing and entanglement dynamics through bottlenecks

3.1 Introduction

The promise of wide-ranging applications of quantum computing has sparked interest in developing scalable quantum architectures. Realistic quantum architectures have constrained interactions, often with fixed connectivity [Bru+19b; Kja+20]. This connectivity is usually specified by a graph G, with vertices V representing qubits, and edges E representing pairs of qubits between which interactions are permitted. On the other hand, quantum algorithms, quantum error correction schemes, and other quantum information-processing protocols often require interactions between arbitrary pairs of qubits. To implement these interactions, quantum compilers effectively simulate all-to-all connectivity by moving quantum information, at the cost of an overhead in the running time. This task, known as quantum routing, aims to make general quantum algorithms feasible on devices with connectivity constraints by efficiently implementing any permutation of qubits on a given architecture graph [CSU19]. Quantum routing can be performed with swap gates using classical techniques [ACG94; Cow+19; Wag+23; CSU19; LDX19], or employing more general quantum operations such as continuous-time Hamiltonian evolution [Bap+21; Bap+23] or unitary gates [Bap+23], possibly assisted by measurement and fast classical feedback [Dev+24; Ros13; HPE19]. Recent work has shown that the worst-case depth overhead of implementing arbitrary quantum circuits while respecting connectivity con-

straints is proportional to the circuit depth of worst-case routing [YZ25], highlighting the importance of lower bounds and fast protocols for this task.

Routing is closely related to the task of entanglement distribution, since by routing entangled qubits, we can distribute entanglement. The rate at which entanglement can be generated between parts of a system constrains our ability to perform many quantum information processing tasks, as entanglement is a critical resource for quantum algorithms [Joz98; JL03; EJ98; BM11; ZZC25], quantum error correction [Bra+25; Ben+96a; KL97], quantum communication [BCD01; GT07; Hor+09], and quantum sensing [DRC17; GLM04; Bol+96]. The fundamental importance of entanglement and its application to a wide range of quantum information tasks has raised many questions and led to a large number of investigations into the dynamics of entanglement [VMV13; Mar+16; CVC05; Cub+03; Bra07; Dür+01; Ben+03].

The fact that routing can be used to distribute entanglement means that bounds on entanglement dynamics can also be used to constrain the time taken to perform routing. Previous work has used this idea to bound routing times [Bap+23] in terms of entanglement rates [VMV13] and entanglement capacities [Mar+16] across system bipartitions. These two major results on entanglement dynamics can be interpreted in terms of bottlenecks, which are of two kinds:

1. Edge bottlenecks: The number of edges across any bipartition of the system constrains the rate at which the bipartitions can be entangled, by the small incremental entangling (SIE) theorem [VMV13]. Thus, in systems with a small number of edges connecting large subsystems, the routing time is large. An example of such system is a d-dimensional array, where the number of edges joining two halves of the system scales as $\Theta(N^{d-1})$, which is asymptotically smaller than the number of vertices on each side ($\Theta(N^d)$). We refer to such

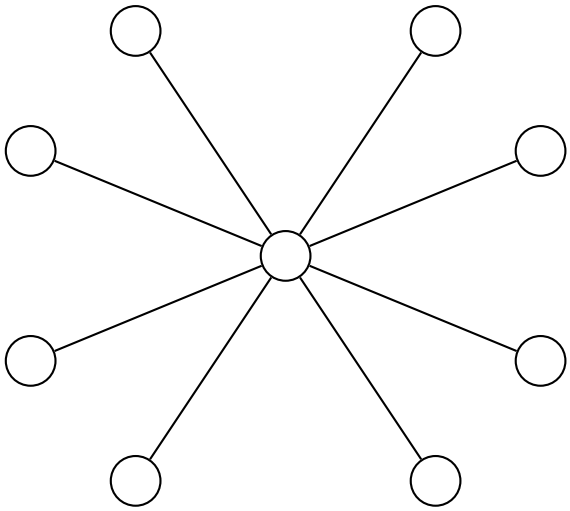


Figure 3.1: The star graph on 9 vertices, S_9.

a set of edges as an edge bottleneck.

2. Vertex bottlenecks: A vertex bottleneck is a set of vertices, typically much smaller than the rest of the system, that constrains the interactions of the rest of the system. An archetypal example of a vertex bottleneck is found in the star graph S_N, shown in Fig. 3.1, which is the complete bipartite graph $K_{1,N-1}$ between one (center) vertex and $N-1$ leaf vertices. Intuitively, to implement interactions between qubits on the leaves, we must make use of the central vertex, which forms a bottleneck. On the other hand, this graph does not have a significant edge bottleneck, due to the large number of edges between the central vertex and the leaves. A general vertex bottleneck is depicted in Fig. 3.2, which represents systems with a tripartite structure. Such systems can be divided into three parts L, C, R such that edges are only present connecting L to C and C to R. The small total entangling property [Mar+16] bounds the entanglement capacity, or total amount of entanglement that can be generated across a partition, in terms of the local dimension of the vertex boundary (sites on one side of a partition that are adjacent to sites on the other side), and therefore constrains

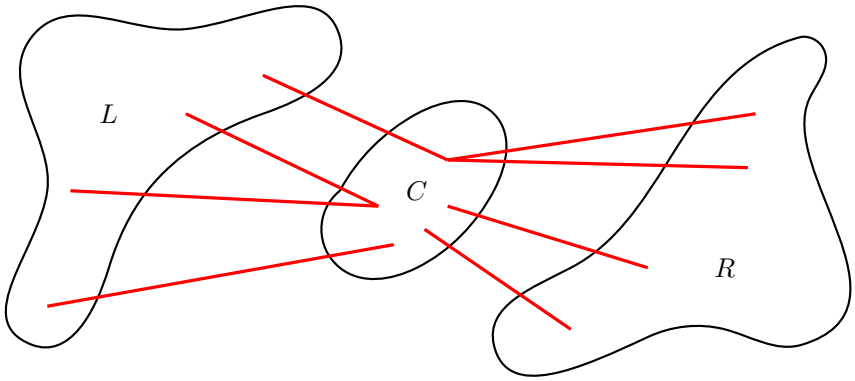


Figure 3.2: A tripartition with vertex bottleneck C. Parts L, C, R contain N_L, N_C, N_R qubits, respectively. By the tripartite connectivity constraint, there are no edges between qubits in L and qubits in R.

the ability to generate entanglement through a vertex bottleneck.

In the gate-based routing model, we make use of 2-qubit gates that act locally, obeying the connectivity constraints of the architecture. A special case of this is classical swap-based routing, where only connectivity-respecting swap gates are permitted. Alternatively, in the more powerful Hamiltonian routing model [Bap+23], we are permitted continuous-time evolution by an architecture-respecting Hamiltonian with norm-bounded interactions. In this work, we primarily investigate piecewise time-independent Hamiltonians. The routing time of a model for a given graph G is defined as the time taken to implement a worst-case permutation on G. In the gate-based model, this corresponds to the minimum circuit depth to implement the permutation, while in the Hamiltonian routing model this corresponds to the minimum evolution time required. These routing models are described in more detail in Section 3.2.

Using bounds on the entangling rate [VMV13], tight bounds can be obtained for routing constrained by edge bottlenecks in both the gate-based and continuous-time evolution routing models. However, for routing through vertex bottlenecks, only entanglement bounds in the gate-

based model [Mar+16] have been found to be tight [Bap+23], following the natural assumption that no two gates can interact with the same qubit at the same time. Attempts at a full treatment of the entanglement dynamics in multipartite systems reveal surprisingly more complexity than in the bipartite case. For example, Cubitt et al. [Cub+03] showed that in a system with a tripartition, entanglement can be generated between the two separated parts without ever entangling the intermediate region with the rest of the system. While this result defies the natural intuition that entanglement must flow between regions of a system, such separable entanglement generation is only possible if the initial state is a special mixed state. For systems starting in a pure state, a notion of entanglement flow between different subsystems was shown by Cubitt et al. [CVC05]. However, this result still provides only a trivial lower bound for routing through vertex bottlenecks.

For the star graph, the gate-based quantum routing time is $\Theta(N)$ [Bap+23]. However, in the Hamiltonian routing model, which has previously only been shown to be constrained by edge bottlenecks, the best known lower bound on the routing time is $\Omega(1)$ (applying the result of Van Acoleyen et al. [VMV13] or a Lieb-Robinson bound [CLY23]). This lower bound intuitively seems loose, since if it were indeed saturable, this would allow for routing in constant time, independent of the system size.

Surprisingly, polynomial speedups for continuous-time evolution over the gate-based model are indeed possible, even in the simple setting of the star graph. For one of our main results, we investigate systems of free (i.e., non-interacting) particles, where the edges of the underlying graph represent two-body hopping terms in the Hamiltonian. In systems of free particles, the gate-based model can be considered to allow architecture-respecting circuits of free-particle gates (i.e., unitaries generated by hopping and on-site interactions). The gate-based routing time

on the star graph for free particles is $\Theta(N)$, by the Small Total Entangling lemma [Mar+16], applied to circuits of free-particle gates. For these systems, we design a fast Hamiltonian routing protocol that takes time $t = O(\sqrt{N})$ for arbitrary permutations, which demonstrates a $\Omega(\sqrt{N})$ speedup over the gate-based model. We further provide a concise proof that this protocol is optimal (for both free fermions and free bosons), showing a lower bound of $t = \Omega(\sqrt{N})$ on the routing time on the star graph for these systems. We note that this lower bound applies even to time-dependent Hamiltonian routing.

Returning to the more standard setting of qubit systems, we significantly tighten the lower bound on the Hamiltonian routing time for the star graph to $\Omega(N^{1/2-\delta})$ for any $\delta > 0$, when using piecewise time-independent Hamiltonians, or time-dependent Hamiltonians subject to a smoothness condition. This addresses an open question of Bapat et al. [Bap+23]. The first part of our proof leverages improved commutator bounds on Hamiltonian simulation for Trotter-Suzuki formulas [Chi+21] for random input states [Zha+22]. This result utilizes the Frobenius norm, allowing us to leverage the fact that vertex-bottlenecked Hamiltonians exhibit tighter commutator scaling in the Frobenius norm compared to the operator norm used in entanglement dynamics bounds [VMV13] or Lieb-Robinson bounds [CLY23]. In contrast to those results, we take advantage of the inherently state-independent task of performing large-scale routing to bound the *average*-case routing time rather than just the worst case, which allows a more precise bound based on the Frobenius norm. The technique behind our proof is inspired by the recent application of "Frobenius light cone" bounds, which demonstrated the hardness of implementing the shift-by-one translation on a ring architecture [YLS24].

This result generalizes to arbitrary vertex bottlenecks, as illustrated in Fig. 3.2. For an arbitrary tripartition of N_L, N_C, N_R qubits (with $N_L \geq N_R \geq N_C$), we show that the routing

time is $\Omega(N_R^{1-\delta}/\sqrt{N_L}N_C)$ for any $\delta > 0$.

Our techniques further allow us to prove an upper bound on the entanglement capacity, or the amount of entanglement that can be generated or distributed through a vertex bottleneck. For a tripartition, we define the entanglement capacity for evolution of a state $|\psi\rangle$ by a given Hamiltonian H for time t as the increase in the amount of bipartite entanglement between the L subsystem and the rest of the system. Existing results [VMV13; CVC05] bound the entanglement capacity as $O(tN_LN_C)$ for any initial state $|\psi\rangle$. For the case where the input state is chosen from a 1-design ensemble, we show an improved upper bound of $O(tN_C\sqrt{N_L}(\sqrt{N_LN_R/N_C})^{\delta})$, for any $\delta > 0$, on the entanglement capacity in systems with a tripartition. By averaging over input states, we obtain a tighter bound that depends on Frobenius norms of commutators between terms of the Hamiltonian, rather than the operator norm of the Hamiltonian as used in Refs. [VMV13; CVC05]. We believe this result could be useful in bounding the time taken for more general quantum information processing tasks beyond routing. For example, our methods may be useful in demonstrating the hardness of implementing specific unitary gates on a quantum processor.

The remainder of the chapter is organized as follows. In Section 3.2, we introduce our models and technical background for the routing problem. In Section 3.3.1, we describe a fast routing protocol for systems of free fermions with star-graph connectivity, which performs routing in time $O(\sqrt{N})$. In Section 3.3.2 we discuss the challenges of converting this to a qubit routing protocol. In Section 3.3.3, we show a concise lower bound of $\Omega(\sqrt{N})$ on the routing time for fermions in systems with a vertex bottleneck, including the star graph, which demonstrates that our protocol in Section 3.3.1 is optimal. In Section 3.4, we show our main result bounding the routing time through vertex bottlenecks in systems of qubits. In Section 3.5, we show how similar techniques provide a lower bound on the average entangling capacity in systems with a tripartition. Finally,

we discuss implications of our results and some open questions in Section 3.6.

3.2 Preliminaries

We let $\|\cdot\|$ denote the spectral/operator norm and $\|\cdot\|_p$ the Schatten p-norm. The (normalized) Frobenius norm of an operator O is

$$\|O\|_{\mathrm{F}} \coloneqq \sqrt{\frac{\mathrm{tr}[O^\dagger O]}{D}}, \tag{3.1}$$

where D is the total Hilbert space dimension. Note that $\|O\|_{\mathrm{F}} = \|O\|_2/\sqrt{D}$. With this normalization, $\|\mathbf{X}\|_{\mathrm{F}} = 1$ for any Pauli string $\mathbf{X}$.

We specify the connectivity of a quantum system with 2-local interactions in the form of a graph $G = (V, E)$, with the vertices V representing qubits/modes and the edges E representing pairs of qubits/modes between which interactions are allowed. In this work, we are mostly concerned with graphs with a vertex bottleneck, as depicted in Fig. 3.2. Routing is the task of implementing arbitrary permutations from $\mathcal{S}_N$, the group of all permutations of N elements, on qubits labeled from 1 to N. For any $p \in \mathcal{S}_N$, where $p(i)$ represents the destination of the i^{th} qubit, we define the permutation unitary U_p, such that, for any product state $|\psi\rangle = |\psi_1\rangle|\psi_2\rangle \dots |\psi_N\rangle$,

$$U_p|\psi_1\rangle|\psi_2\rangle \dots |\psi_N\rangle = |\psi_{p(1)}\rangle|\psi_{p(2)}\rangle \dots |\psi_{p(N)}\rangle. \tag{3.2}$$

3.2.1 Hamiltonian routing

In the Hamiltonian routing model, we consider continuous-time evolution by a Hamiltonian H that respects the connectivity constraints. Such an architecture-respecting evolution has no interaction terms between sites (qubits) that are not connected by an edge in the underlying graph G. Thus, any valid Hamiltonian can be written as

$$H = \sum_{(i,j)\in E} h_{ij} + \sum_{i\in V} h_i, \tag{3.3}$$

where h_{ij} is only supported on qubits i, j (i.e., acts as identity on all other qubits). We further impose the constraint $\|h_{ij}\| \leq 1$ in order to ensure that the timescale of two-site interactions is comparable to the time required for a 2-qubit gate in the gate-/swap-based routing models. Unitary evolution by a Hamiltonian H is given by the time-evolution operator

$$U(H,t) \coloneqq \mathcal{T} e^{-\mathrm{i}\int_0^t H(t')\,\mathrm{d}t'}, \tag{3.4}$$

where $\mathcal{T}$ is the time-ordering operator. When H is time-independent, $U(H,t) = e^{-\mathrm{i}Ht}$.

The *Hamiltonian routing time* is the minimum time to implement a given permutation in this model, defined as

$$\mathrm{hrt}(G,p) \coloneqq \min_H \{t \text{ s.t. } U(H,t) = U_p\} \tag{3.5}$$

for any $p \in \mathcal{S}_N$. We define the worst-case Hamiltonian routing time for a graph G as

$$\mathrm{hrt}(G) \coloneqq \max_{p\in\mathcal{S}_N} \mathrm{hrt}(G,p). \tag{3.6}$$

In this work, we primarily consider time-independent or piecewise time-independent H.

Definition 3.1 (Piecewise time-independent Hamiltonian)**.** H is said to be piecewise time-independent with minimum segment width Δ if it can be written as

$$H = \sum_{i=0}^{k-1} H_i \cdot \mathbb{I}_{t\in[t_i,t_{i+1})}, \tag{3.7}$$

where $t_k = t$, H_i is time-independent, $\mathbb{I}_{t\in[t_i,t_{i+1})}$ is an indicator function for $t \in [t_i, t_{i+1})$, and $\forall\, i, t_{i+1} - t_i \geq \Delta$.

Similarly to the time-dependent case, we define the Hamiltonian routing time for piecewise time-independent Hamiltonians with minimum segment width Δ as

$$\mathrm{hrt}_\Delta(G, p) \coloneqq \min_H \{t \text{ s.t. } U(H, t) = U_p\} \tag{3.8}$$

for any $p \in \mathcal{S}_N$. We define the worst-case Hamiltonian routing time for piecewise time-independent Hamiltonians with minimum segment width Δ for a graph G as

$$\mathrm{hrt}_\Delta(G) \coloneqq \max_{p\in\mathcal{S}_N} \mathrm{hrt}_\Delta(G, p). \tag{3.9}$$

3.2.2 Gate-based routing

The gate-based routing model is a variant of the Hamiltonian routing model where we are restricted to performing 1-qubit and 2-qubit gates that respect the connectivity constraints (i.e., no gates are permitted between qubits not connected by an edge). Gates on disjoint pairs of vertices may be applied simultaneously. In this model, the routing time for a given permutation is the

minimum depth to implement a given permutation p by an architecture-respecting circuit,

$$\mathrm{rt}(G, p) \coloneqq \min_{U} \mathrm{depth}(U) \text{ s.t. } U = U_p. \tag{3.10}$$

Similar to the Hamiltonian routing model, we also define the *worst-case routing time* for a graph G as

$$\mathrm{rt}(G) \coloneqq \max_{p \in \mathcal{S}_N} \mathrm{rt}(G, p). \tag{3.11}$$

The gate-based model generalizes classical routing, which is restricted to circuits composed of architecture-respecting swap gates [ACG94].

3.2.3 Free-particle routing

An alternative type of Hamiltonian routing is free-particle routing (of bosons or fermions).

In a fermionic model, each vertex of G represents a local fermionic mode, which can be either empty ($|0\rangle$) or occupied by a fermion ($|1\rangle$). Fermionic states and interactions are represented in terms of creation and annihilation operators on each mode $(c_j^\dagger, c_j)$ [BK02].

The annihilation operator acts as follows:

$$\begin{aligned} &c_j |n_0, n_1, \ldots, n_{j-1}, 1, n_{j+1}, \ldots, n_N\rangle \\ &\qquad = (-1)^{\sum_{k=0}^{j-1} n_k} |n_0, n_1, \ldots, 0_j, \ldots, n_N\rangle, \end{aligned} \tag{3.12a}$$

$$c_j |n_0, n_1, \ldots, n_{j-1}, 0, n_{j+1}, \ldots, n_N\rangle = 0. \tag{3.12b}$$

These operators obey the fermionic anticommutation relations

$$\{c_j, c_k\} = \{c_j^\dagger, c_k^\dagger\} = 0,\ \{c_j, c_k^\dagger\} = \delta_{j,k}. \tag{3.13}$$

We also consider routing in systems of free (non-interacting) bosons. Similarly to the fermionic case, states are characterized as Fock states, i.e., in the occupancy number basis $|n_i\rangle$ where n_i is the occupancy of the i^{th} mode (situated on site i). The bosonic creation and annihilation operators are $(a_j^\dagger, a_j)$. The annihilation operator acts as follows:

$$\begin{aligned} &a_j|n_0, n_1, \ldots, n_{j-1}, n_j, n_{j+1}, \ldots, n_N\rangle \\ &\quad = \sqrt{n_j}|n_0, n_1, \ldots, n_j - 1, \ldots, n_N\rangle. \end{aligned} \tag{3.14}$$

These operators obey the bosonic commutation relations

$$[a_j, a_k] = [a_j^\dagger, a_k^\dagger] = 0,\ [a_j, a_k^\dagger] = \delta_{j,k}. \tag{3.15}$$

Any fermionic or bosonic Hamiltonian can be written as a sum of products of creation and annihilation operators. An architecture-respecting fermionic/bosonic Hamiltonian can be written as a sum of products of these operators on sites that are connected by an edge in G, and have bounded coefficients. In a free (non-interacting) model, the Hamiltonian only contains hopping and on-site terms (in Section 3.6, we comment on the possibility of allowing for pairing). In

particular, a free-particle Hamiltonian can be written as

$$H = \sum_{i \in V} h_i c_i^\dagger c_i + \sum_{(i,j) \in E} \left(h_{ij} c_i^\dagger c_j + \text{h.c.} \right), \tag{3.16}$$

where $|h_{ij}| \leq 1$, h_i, h_{ij} may have arbitrary time dependence, and c_i are all fermionic or all bosonic.

A permutation $p \in \mathcal{S}_N$ maps each mode on site i to site $p(i)$. Therefore, U_p maps the set of annihiliation operators $\{c_1, c_2, \dots\}$ to $\{c_{p(1)}, c_{p(2)}, \dots\}$. For particles evolving by a free Hamiltonian with modes b_i, the time-evolved creation and annihilation operators at a given time t can be written as $b_i(t) = \sum_j A_{ij}(t) b_j(0)$ and $b_i^\dagger(t) = \sum_j A_{ij}^*(t) b_j^\dagger(0)$ for some $A_{ij}(t) \in \mathbb{C}$. The dynamics of $A_{ij}(t)$ are the same for both bosons and fermions (and for a single particle). Hence, any protocol that performs routing for free fermions yields an identical routing protocol for free bosons and vice versa.

We define the free-particle routing time as the minimum time to implement a given permutation using an architecture-respecting free-particle Hamiltonian:

$$\text{hrt}_F(G, p) \coloneqq \min_H [t \text{ s.t. } U(H, t) = U_p], \tag{3.17}$$

where $U(H, t)$ is the time-evolution operator by H for time t. We define the *worst-case free-particle routing time* for a graph G as

$$\text{hrt}_F(G) \coloneqq \max_{p \in \mathcal{S}_N} \text{rt}_F(G, p). \tag{3.18}$$

Since the dynamics of the creation and annihilation operators are identical for bosons and

fermions, the routing time does not depend on the particle type.

One can also define a gate-based routing model for free particles. Allowed two-mode gates between mode i and mode j for $(i,j) \in E$ are the unitaries generated by $h_i c_i^\dagger c_i + h_j c_j^\dagger c_j + (h_{ij} c_i^\dagger c_j + \text{h.c.})$ for time-dependent h_i, h_j, and h_{ij}. As in the case of qubits, gates on disjoint pairs of vertices may be applied simultaneously, and application of a single layer of allowed simultaneous gates takes depth (or time) 1. We define the gate-based free-particle routing time $\mathrm{rt}_F(G,p)$ as the minimum depth to implement a given permutation p using an architecture-respecting free-particle circuit (i.e., a circuit composed of free-particle gates). We define the *worst-case* gate-based free-particle routing time for a graph G as $\mathrm{rt}_F(G) \coloneqq \max_{p\in\mathcal{S}_N} \mathrm{rt}_F(G,p)$.

3.2.4 Transpositions

Definition 3.2 (Graph Tripartition)**.** A tripartition of a graph G is a partitioning of the vertices of G into three sets L, C, R with N_L, N_C, N_R vertices, respectively, such that there are no edges connecting vertices in L to vertices in R. Without loss of generality, L is taken to be the larger of L, R, i.e., $N_L \geq N_R$.

A graph tripartition is depicted in Fig. 3.2.

Definition 3.3 (Vertex bottleneck)**.** In a graph with a tripartition, C is a vertex bottleneck if $N_C \leq N_R$.

Throughout this chapter, we consider tripartitions with a vertex bottleneck, i.e., we assume everywhere that $N_L \geq N_R \geq N_C$. We focus on permutations that perform the maximum possible number N_R of pairs of disjoint swaps between sites through a general vertex bottleneck (Fig. 3.2). On the star graph S_{N+1}, such a permutation can be obtained by dividing the leaves into two sets

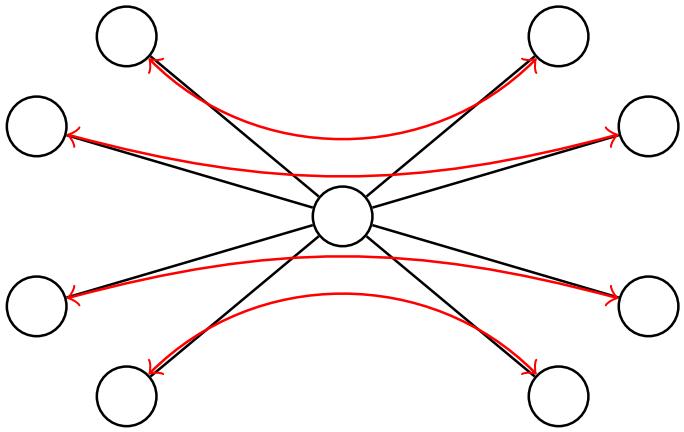


Figure 3.3: A global transposition permutation on the star graph. The black lines denote the connectivity constraints. The red lines with arrows indicate pairs of qubits to be swapped.

L, R of $N/2$ qubits each, and swapping each qubit of L with a corresponding qubit of R. We call this a global transposition, and it is depicted in Fig. 3.3.

Our investigation of such permutations is motivated by the fact that they can be used at most twice to implement an arbitrary permutation. Any cycle, or set of non-overlapping cycles, can be performed using two global transpositions. Since any permutation can be decomposed into a product of non-overlapping cyclic permutations, arbitrary permutations can be performed using two global transpositions. For example, the cyclic permutation $1 \to 2, 2 \to 3, \ldots, 2n \to 1$, which we denote by $(1\ 2 \ldots 2n)$, can be decomposed as a product of two global transpositions:

$$\begin{aligned}(1\ 2 \ldots 2n) = (2\ 2n)(3\ 2n-1)\cdots(n\ n+2) \\ \times (1\ 2n)(2\ 2n-1)\cdots(n\ n+1),\end{aligned} \tag{3.19}$$

where the two lines on the right-hand side each consist of non-overlapping pairwise transpositions. This is illustrated in Fig. 3.4. On the star graph, any permutation can be decomposed into

a permutation that acts only on the leaves, followed by a single final swap with the center node. Hence, the time taken for arbitrary permutations on the star graph is predominantly determined by the time to do a global transposition.

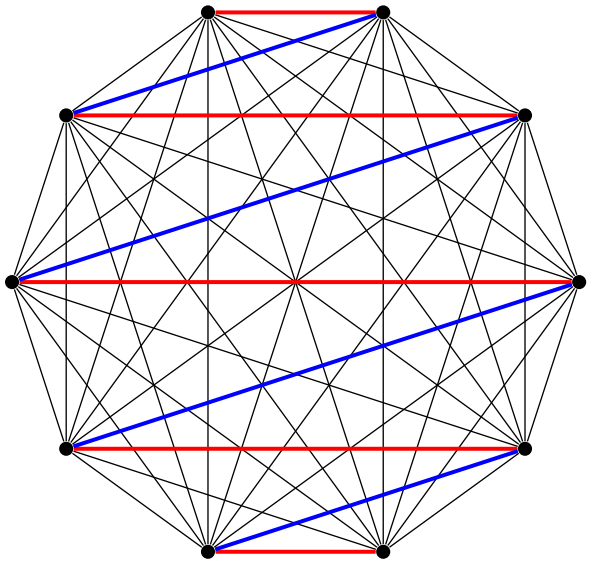
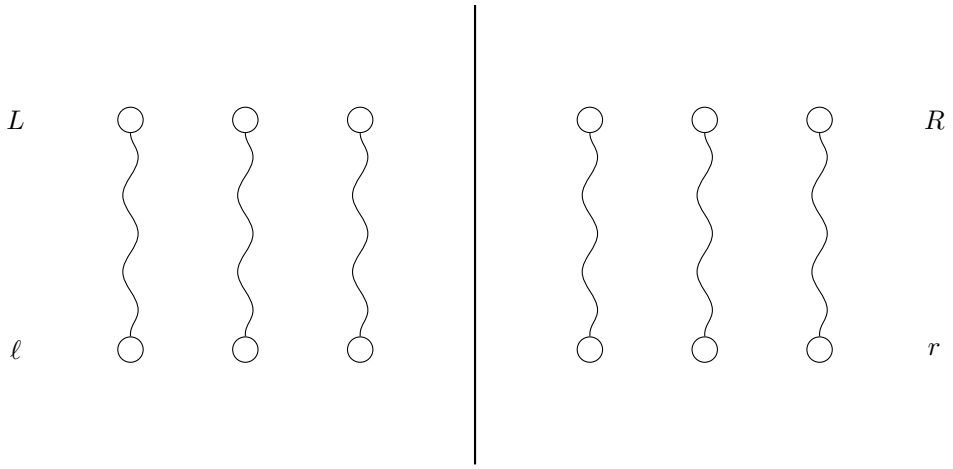


Figure 3.4: Any cyclic permutation can be decomposed into two stages of transpositions. For example, a clockwise cycle of all vertices can be achieved by a global transposition between pairs of sites connected by blue edges followed by a global transposition for red edges. To perform an anti-clockwise cycle, simply transpose along the red edges and then the blue edges.

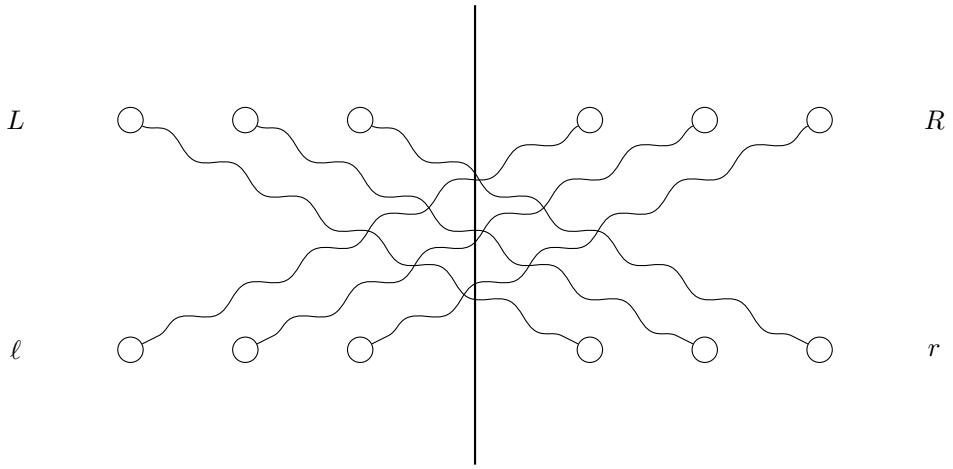
3.2.5 Routing and entanglement distribution

Routing can be used to distribute entanglement, as depicted in Fig. 3.5. This fact allows us to lower bound the routing time using bounds on the time to generate entanglement. We make use of two such bounds.

The small incremental entangling theorem, whose conjecture is attributed to Kitaev in Ref. [Bra07] and which is proved in Ref. [VMV13], bounds the rate at which entanglement can be generated between parts of a system in terms of the local dimension and the operator norm of the Hamiltonian coupling the parts. We restate this result here:



(a) Initial state with local bell pairs



(b) Bell pairs distributed after routing

Figure 3.5: Routing enables entanglement distribution. (a) The system has two parts L, R that are initially entangled with hidden auxiliary systems l, r, respectively. (b) On swapping every qubit in L with the corresponding qubit in R, the L system becomes entangled with r, and R becomes entangled with l, so we have generated 6 ebits of entanglement across the cut dividing L and R.

Lemma 3.4 (Small Incremental Entangling (SIE))**.** *Given a Hamiltonian* $H = H_{Aa} + H_{Bb} + H_{AB}$ *that acts on a system consisting of subsystems* a, A, B, b *(Fig. 3.6), for any state* ρ*,*

$$\frac{\mathrm{d}S_{aA}(\rho)}{\mathrm{d}t} \leq c\|H_{AB}\| \log(d), \tag{3.20}$$

where $S_{Aa}(\rho) = -\operatorname{tr}[\rho_A \log \rho_A]$ *is the entanglement entropy of* ρ *across the* aA, Bb *bipartition,* c *is some positive constant (*$c = 2$ *in [Aud14]), and* $d = \min\{\dim(A), \dim(B)\}$.

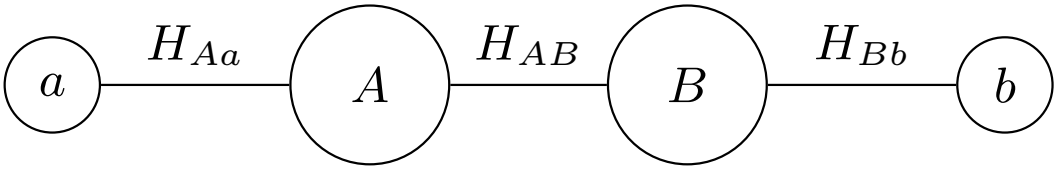


Figure 3.6: Partitioning of system for SIE.

We also make use of the small total entangling property [Mar+16; Bap+23], which bounds the total amount of entanglement between parts of a system that can be generated by any unitary evolution, in terms of the local dimension of the parts.

Lemma 3.5 (Small Total Entangling (STE) (Proposition 2 of Ref. [Mar+16]))**.** *In a system consisting of subsystems* a, A, B, b*, for any unitary acting only on* A *and* B*, the maximum change in the entanglement entropy of the* Aa *subsystem is bounded as*

$$\Delta S_{Aa}(\rho) \leq 2 \log(\min\{\dim(A), \dim(B)\}). \tag{3.21}$$

In systems of free particles, we quantify bipartite entanglement by the *mode entanglement* of the underlying Fock state [Shi03; Zan02], which is the von Neumann entropy obtained by tracing out a subset of the modes.

Definition 3.6 (Mode entanglement)**.** For a state of identical particles ρ (i.e., a density matrix in the occupation number basis) with modes partitioned into X and $\bar{X}$, the mode entanglement is

$$S_X(\rho) = -\operatorname{tr}[\rho_X \log \rho_X]. \tag{3.22}$$

3.3 Routing free particles

3.3.1 An optimal protocol for routing free particles

Here we consider the problem of routing free particles on a star graph. We give a protocol that can route N free fermions on the star graph in time $\sqrt{N}$. Like the W-state-based routing protocol on the vertex barbell graph (Fig. 3.9) of Bapat et al. [Bap+23], this makes use of states spread over multiple sites (here, the leaves) and transfers them one-by-one through the central vertex.

Theorem 3.7. $\operatorname{hrt}_F(S_N) = O(\sqrt{N})$.

Proof. We consider the star graph S_{2N+1}, i.e., a star graph with an even number of leaves. To see that this is sufficient, consider any permutation p on a star graph with an odd number of leaves. We can select some leaf i, and perform a modified permutation p' which is identical to p except when acting on sites i and $p^{-1}(i)$. The permutation p' instead maps $p^{-1}(i)$ to $p(i)$ and leaves i unchanged, and is therefore a permutation on an even number of leaves. By first performing p', and then swapping qubits i and $p(i)$, we can perform the permutation p. Hence it suffices to consider a star graph with an odd number of vertices (i.e., an even number of leaves), since for any permutation p on a star graph with an odd number of leaves $2N + 1$, the routing time is only

a constant (3) greater than the routing time for a corresponding permutation on a subgraph with an even number of leaves $2N$.

As explained in Section 3.2, an arbitrary permutation can be implemented by application of at most two global transpositions. For any global transposition, the leaf vertices can be divided into two sets L (left) and R (right) with N leaves on each side, such that the transposition swaps each qubit in L with the corresponding qubit in R. Let the pairs of creation and annihilation operators on each left (right) leaf i be $l_i^\dagger$ and l_i ($r_i^\dagger$ and r_i), respectively. On the center, we denote them as $c^\dagger$ and c.

The left Fourier modes are $f_{l,0}, \ldots, f_{l,N-1}$, where

$$f_{l,k} \coloneqq \frac{1}{\sqrt{N}} \sum_j e^{-2\pi \mathrm{i} kj/N} l_j. \tag{3.23}$$

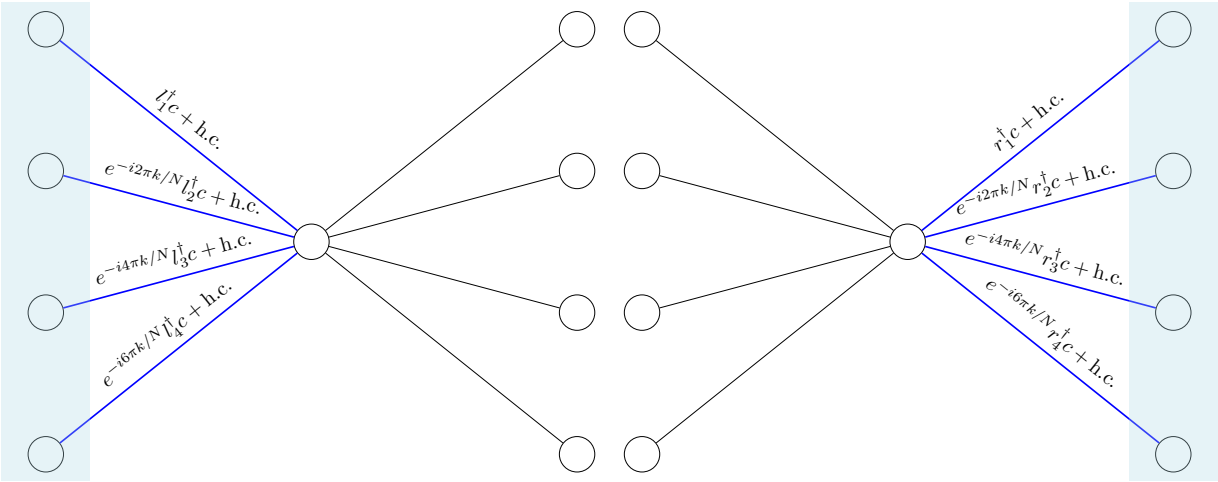
Likewise, we denote the right Fourier modes as $f_{r,0}, \ldots, f_{r,N-1}$.

We divide our protocol into N time steps, numbered 0 to $N-1$. At the kth time step, we perform the following operations:

- First, turn on coupling $H_{L,k} = \sum_j e^{-2\pi \mathrm{i} jk} c^\dagger l_j + \mathrm{h.c.}$ for time $\frac{\pi}{2\sqrt{N}}$. This Hamiltonian can be equivalently written as $\sqrt{N}(c^\dagger f_{l,k} + \mathrm{h.c.})$ This swaps the center mode with the kth Fourier mode on the left. This is illustrated in Fig. 3.7a.

- Repeat the same step but with the right leaves using the coupling $H_{R,k} = \sum_j e^{-2\pi \mathrm{i} jk} c^\dagger r_j + \mathrm{h.c.}$ This interaction swaps $f_{l,k}$ (which sits at the center following the first step) with $f_{r,k}$ such that $f_{l,k}$ is now on the *right* leaves, $f_{r,k}$ is on the center, and c has been moved to the kth left Fourier mode. This is illustrated in Fig. 3.7b.

- Finally, perform the coupling $H_{L,k} = \sum_j e^{-2\pi ijk} c^\dagger l_j + \text{h.c.}$ with the left leaves again to swap c and $f_{r,k}$. We have now effectively swapped Fourier modes $f_{l,k}$ and $f_{r,k}$, with c remaining on the center. This is illustrated in Fig. 3.7a.



(a) Left swap with center

(b) Right swap with center

Figure 3.7: Swapping the k^{th} Fourier modes on the left and right sets of leaves.

Since both Fourier modes and spatial modes form orthonormal bases, swapping all of the corresponding Fourier modes between the left and the right vertices is equivalent to swapping each of the corresponding spatial modes. Since each step of this protocol swapping one Fourier mode at a time takes time $O(\frac{1}{\sqrt{N}})$, the whole protocol takes time $O(\sqrt{N})$. As discussed in Section 3.2, the dynamics of free bosonic and free fermionic creation and annihilation operators are identical. Hence this protocol performs routing for both bosons and fermions. □

3.3.2 Routing qubits

A natural question is whether a similar protocol exists for qubits. Unfortunately, simply converting the protocol above to a qubit protocol does not work, as we now explain.

First, consider the fermionic version of this protocol. Applying the Jordan-Wigner [JW28] transform to convert the fermionic operators to qubit operators yields non-local terms, so the generated qubit Hamiltonian does not respect the architecture's geometry and hence does not produce a valid protocol. Even if we permitted ancillary qubits, using an approach such as the Verstraete-Cirac encoding [VC05] to ensure that we respected locality would introduce a large number of ancillas ($\Theta(N)$) on the central qubit due to the large number of non-commuting terms in the Hamiltonian.

Now we consider whether a bosonic protocol can be converted to a qubit protocol. One approach is to consider the bosonic protocol restricted to an initial state where each mode has occupancy at most 1, and replace the bosonic creation and annihilation operators with spin-1/2 raising and lowering operators, respectively. Another similar strategy to convert a bosonic protocol to a qubit routing protocol is by truncating the bosonic operators to a low (constant) occupancy subspace by replacing the bosonic operators with qudit operators, which may yield an approximate protocol for qubits with ancillas.

Unfortunately, such an approach is infeasible. Consider a bosonic system starting in the uniform superposition of all bitstrings of length N with Hamming weight $N/2$, i.e., the weight $N/2$ Dicke state $|W^N_{N/2}\rangle$. Since every mode has occupancy at most 1, this is a valid initial state for a qubit system as well. As before, we denote the left Fourier modes as $f_{l,0}, \ldots, f_{l,N-1}$, where

$$f_{l,k} \coloneqq \frac{1}{\sqrt{N}} \sum_j e^{-2\pi \mathrm{i} kj/N} l_j. \tag{3.24}$$

Likewise, we denote the right Fourier modes as $f_{r,0}, \ldots, f_{r,N-1}$. Then, a simple calculation shows that the expected occupancy of the zeroth Fourier mode $f_{l,0}$ is $\langle W^N_{N/2}|f^\dagger_{l,0} f_{l,0}|W^N_{N/2}\rangle = \frac{\sqrt{N+2}}{4}$.

Hence, if we were to truncate the protocol using qudit operators, it would require qudits of local dimension $\Omega(\sqrt{N})$. Such an approach therefore cannot yield a qubit routing protocol.

Though we have ruled out some obvious approaches, it may still be possible to produce a similar protocol for qubits (or to rule one out entirely), which we believe are interesting directions for future work.

3.3.3 Lower bound on the free-particle routing time

In this section, we give lower bounds on the time to route N non-interacting particles (fermions or bosons) through a vertex bottleneck.

First, we show a lower bound of $\Omega(N_R/N_C)$ on the gate-based routing model for graphs with a vertex bottleneck. On the star graph, this bound yields $\mathrm{rt}_F(S_N) = \Omega(N)$. This implies that our Hamiltonian-based protocol in Section 3.3.1 obtains a quadratic speedup over gate-based routing.

Then, we show a lower bound on the Hamiltonian routing time $\mathrm{hrt}_F(G)$ for any graph G with a vertex bottleneck. In particular, our proof implies that the free-particle routing time on the star graph is lower-bounded by $\mathrm{rt}_F(S_N) = \Omega(\sqrt{N})$, which implies that the free-particle routing protocol given in Section 3.3.1 is optimal.

To prove these routing-time lower bounds, we first lower bound the time to generate $\Theta(N)$ mode entanglement (Definition 3.6) between two halves of a free-particle system. Analogously to the discussion in Section 3.2.5, the ability of a system to route N non-interacting particles across a partition in time t can be reduced to the problem of generating $\Theta(N)$ mode entanglement in that time—for example, when N modes, each of which is entangled with an ancillary mode as

the state $\frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$, are routed across a partition, the amount of entanglement generated across the partition is $\Theta(N)$. As such, a lower bound on the time to generate $\Theta(N)$ entanglement across a vertex bottleneck can be used to lower bound the time to perform routing.

To fix the model more precisely, we assume the system is defined on an architecture graph with a tripartition $V = L \cup C \cup R$ (cf. Fig. 3.2), where each vertex represents a mode and each edge indicates pairs of modes between which we are allowed to apply a gate. We now show a lower bound on the gate-based routing time for free particles.

Theorem 3.8. *For any graph G with a vertex bottleneck,* $\mathrm{rt}_F(G) = \Omega(N_R/N_C)$.

Proof. We first consider routing in a system of free fermions. As discussed in Section 3.3, due to their identical operator dynamics, any lower bound for routing free fermions also applies to routing free bosons. Consider an initial state $\rho_{\text{in}} = \frac{1}{2^{N_R}}\mathbb{I}_R \otimes |0\rangle^{N_C+N_L}$. This can be considered as a state where each mode in R is maximally entangled with some reference system that has been traced out. The initial von Neumann entropy of the subsystem R is thus $S_R = N_R$. On routing every mode from R to L, all the entanglement (or entropy) is transferred to L, and $S_R = 0$. Following a similar argument to one presented in Bapat et al. [Bap+23], by a straightforward application of STE (Lemma 3.5) to free fermions, the maximum change in the von Neumann entropy of R when applying a circuit of depth d is

$$\Delta S_R \leq 2dN_C. \tag{3.25}$$

On the other hand, to route every mode in R to L requires a total change in the von Neumann entropy of R of $|\Delta S_R| = N_R$. Hence, the minimum circuit depth to perform routing is $\mathrm{rt}_F(G) = \Omega(N_R/N_C)$. □

Applied to the star graph, this implies a bound of $\mathrm{rt}_F(S_N) = \Omega(N)$, and hence our Hamiltonian routing protocol from Section 3.3.1 obtains a quadratic speedup over gate-based routing.

We now show a lower bound on Hamiltonian free-particle routing in graphs with a vertex bottleneck. As before, each vertex represents a mode and each edge corresponds to a possible two-body "hopping" term in the Hamiltonian. Let the particles evolve under the following non-interacting (time-dependent) Hamiltonian:

$$\begin{aligned} H(t) &= H^{(LL)} + H^{(CC)} + H^{(LC)} \\ &= \sum_{i,j\in L} H_{ij}^{(LL)}(t) l_i^\dagger l_j + \sum_{i,j\in C} H_{ij}^{(CC)}(t) c_i^\dagger c_j \\ &+ \sum_{\substack{i\in L \\ j\in C}} \left[H_{ij}^{(LC)}(t) l_i^\dagger c_j + \mathrm{h.c.} \right] \\ &+ \sum_{\substack{i\in R \\ j\in C}} \left[H_{ij}^{(RC)}(t) r_i^\dagger c_j + \mathrm{h.c.} \right] + \sum_{i,j\in R} H_{ij}^{(RR)}(t) r_i^\dagger r_j, \end{aligned} \tag{3.26}$$

where the $l_i^\dagger$ (l_i) are the creation (annihilation) operators on L, $c_i^\dagger$ (c_i) are the corresponding operators on C, and $r_i^\dagger$ (r_i) are the corresponding operators on R.

As before, we assume that Hamiltonian terms coupling L and C are bounded by $\|H_{ij}^{(LC)}(t)\| \leq 1$, for all times t.

The following lemma bounds the entangling rate in a free fermionic system. We use this to bound the free fermionic Hamiltonian routing time, and therefore the Hamiltonian routing time for free bosons as well.

Lemma 3.9. *For a free fermionic system defined on a graph with a vertex bottleneck evolving under the free-particle Hamiltonian defined in Eq. (3.26), the entangling rate between subsystems*

L and $C \cup R$ is bounded as

$$\frac{\mathrm{d}S_L(\rho(t))}{\mathrm{d}t} \leq 4cN_C\sqrt{N_L} \tag{3.27}$$

Proof. The proof of the theorem follows from applying the Small Incremental Entangling (SIE) theorem (Lemma 3.4) to a free fermionic Hamiltonian acting on a bipartition of the system $(L, C \cup R)$.

By SIE, we are only concerned with $H^{(LC)} = \sum_{\substack{i\in L\\ j\in C}} \left[H_{ij}^{(LC)}(t) l_i^\dagger c_j + \text{h.c.}\right]$. Consider the single normalized mode

$$\tilde{l}_j^\dagger(t) \coloneqq \frac{1}{\sqrt{\sum_{i\in L}\left|H_{ij}^{(LC)}(t)\right|^2}} \sum_{i\in L} H_{ij}^{(LC)}(t) l_i^\dagger. \tag{3.28}$$

Then

$$H^{(LC)} = \sum_{j\in C} h_j(t) \tag{3.29}$$

where $h_j(t) = \sqrt{\sum_{i\in L}\left|H_{ij}^{(LC)}(t)\right|^2}\left[\tilde{l}_j^\dagger(t) c_j + \text{h.c.}\right]$.

Taking the same approach as in the proof of the area law for quasi-adiabatic continuation in [VMV13], we can bound the entangling rate as

$$\begin{aligned}
\frac{\mathrm{d}S_L(\rho(t))}{\mathrm{d}t} &= -\mathrm{i}\,\mathrm{tr}[H^{(LC)}(t)[\rho(t), \log(\rho_L(t)\otimes \mathbf{1}_C)]] \\
&= -\mathrm{i}\sum_{j\in C} \mathrm{tr}[h_j(t)[\rho(t), \log(\rho_L(t)\otimes \mathbf{1}_C)]] \\
&\leq 2c\sum_{j\in C} \|h_j(t)\|,
\end{aligned} \tag{3.30}$$

where $\rho_L(t) = \mathrm{tr}_C[\rho(t)]$, and c is the constant in Lemma 3.4. In the third line, we applied the SIE

theorem to bound the contribution of each $h_j(t)$ term separately (cf. Eq. (18) in [VMV13]).

To finish the proof, we bound the norm of $h_j(t)$ as follows:

$$\begin{aligned} \|h_j(t)\| &= \left\| \sqrt{\sum_{i\in L} \left| H_{ij}^{(LC)}(t) \right|^2} \left[\tilde{l}_j^\dagger(t) c_j + \text{h.c.} \right] \right\| \\ &\leq 2\sqrt{N_L}, \end{aligned} \tag{3.31}$$

where we have used the fact that, for fermions, $\|\tilde{l}_j^\dagger(t) c_j\| \leq 1$. Plugging this bound on $\|h_j(t)\|$ into Eq. (3.30) yields

$$\frac{\mathrm{d}S_L(\rho(t))}{\mathrm{d}t} \leq 4cN_C\sqrt{N_L}, \tag{3.32}$$

as claimed. □

Theorem 3.10. *For a system defined on a graph with a vertex bottleneck evolving under the free-particle Hamiltonian defined in Eq. (3.26),*

$$\mathrm{hrt}_F(G) = \Omega\left(\frac{N_R}{N_C\sqrt{N_L}}\right). \tag{3.33}$$

Proof. We start with the case of free fermions. By Lemma 3.9, the entangling rate satisfies $\frac{\mathrm{d}S_L(\rho(t))}{\mathrm{d}t} \leq 4cN_C\sqrt{N_L}$. Hence the time to generate $S_L(\rho(t)) = \Theta(N_R)$ ebits of entanglement between subsystems L and $R \cup C$ is lower bounded by

$$t = \Omega\left(\frac{N_R}{\sqrt{N_L}N_C}\right). \tag{3.34}$$

If every mode in R is initially maximally entangled with an ancillary mode, then routing every

mode from R to L requires producing N_R bits of mode entanglement between L and the rest of the system. Hence, the routing time for free fermions (and thus for free bosons as well) is lower bounded as $\Omega\left(\frac{N_R}{\sqrt{N_L}N_C}\right)$ as claimed. □

We now apply Theorem 3.10 to the case of the star graph. Setting $N_L = N$ and $N_C = 1$ yields a lower bound of $t = \Omega(\sqrt{N})$ on the free-particle routing time, which matches the upper bound given by the free-fermion routing protocol in Section 3.3.1. As such, for free particles, the optimal routing time on the star graph is $\mathrm{hrt}_F(S_N) = \Theta(\sqrt{N})$.

3.4 Lower bounds on qubit routing through vertex bottlenecks

We now return to the more common setting of qubits rather than free particles. An architecture-respecting Hamiltonian for a graph with a tripartition can be decomposed as

$$H = H_L + H_C + H_R, \tag{3.35}$$

where H_L consists of local terms in L, terms coupling different sites in L, and coupling sites in L to sites in C; H_C consists of terms local in C or coupling different sites in C; and H_R consists of terms local in R, coupling different sites in R, and coupling sites in R to sites in C. These terms

can be further expanded in terms of Pauli operators $\{X^0, X^1, X^2, X^3\}$:

$$\begin{aligned} H_L &= \sum_{l_1}^{N_L} h_{l_1}^{(L)\alpha_1} X_{l_1}^{\alpha_1} + \sum_{l_1<l_2}^{N_L,N_L} h_{l_1,l_2}^{(LL)\alpha_1\alpha_2} X_{l_1}^{\alpha_1} X_{l_2}^{\alpha_2} + \sum_{l_1,c_1}^{N_L,N_C} h_{l_1,c_1}^{(LC)\alpha_1\gamma_1} X_{l_1}^{\alpha_1} X_{c_1}^{\gamma_1}, \\ H_C &= \sum_{c_1}^{N_C} h_{c_1}^{(C)\gamma_1} X_{c_1}^{\gamma_1} + \sum_{c_1<c_2}^{N_C,N_C} h_{c_1,c_2}^{(CC)\gamma_1\gamma_2} X_{c_1}^{\gamma_1} X_{c_2}^{\gamma_2}, \\ H_R &= \sum_{r_1}^{N_R} h_{r_1}^{(R)\beta_1} X_{r_1}^{\beta_1} + \sum_{r_1<r_2}^{N_R,N_R} h_{r_1,r_2}^{(RR)\beta_1\beta_2} X_{r_1}^{\beta_1} X_{r_2}^{\beta_2} + \sum_{r_1,c_1}^{N_R,N_C} h_{r_1,c_1}^{(CR)\beta_1\gamma_1} X_{r_1}^{\beta_1} X_{c_1}^{\gamma_1}, \end{aligned} \tag{3.36a}$$

where we use the Einstein summation convention (i.e. summing over repeated indices) for $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2 \in \{1, 2, 3\}$. We assume the coefficients satisfy

$$\begin{aligned} &|h_{l_1}^{(L)\alpha_1}|, |h_{c_1}^{(C)\gamma_1}|, |h_{r_1}^{(R)\beta_1}| \leq \sqrt{N}, \\ &|h_{l_1,l_2}^{(LL)\alpha_1\alpha_2}|, |h_{l_1,c_1}^{(LC)\alpha_1\gamma_1}| \leq 1, \\ &|h_{c_1,c_2}^{(CC)\gamma_1\gamma_2}|, |h_{r_1,r_2}^{(RR)\beta_1\beta_2}|, |h_{r_1,c_1}^{(CR)\beta_1\gamma_1}| \leq 1 \end{aligned} \tag{3.37}$$

with $N = N_L + N_R + N_C$ and $N_C \leq N_L, N_R$. It is natural to assume even the single-site terms must be bounded by 1. However, some routing models may allow for faster local fields [Bap+23]. In this work, we assume a bound of $\sqrt{N}$ for technical reasons to retain our bound on the Frobenius commutator norm in Lemma B.1.1. Since our graph has a bottleneck, $N_L \geq N_R \geq N_C$. These coefficients may be time-dependent such that H is piecewise time-independent (recall Definition 3.1) with minimum segment width $\propto 1/\sqrt{N_L}$ (this value is due to the limitations of our techniques, as expressed by Eq. (3.43)). Our Hamiltonian can alternatively be written as $H = H_{LC} + H_R$, where $H_{LC} = H_L + H_C$.

Our main result is a lower bound on the Hamiltonian routing time for any graph with a

vertex bottleneck.

Theorem 3.11. *For any graph G with a tripartition into L, C, R with N_L, N_C, N_R vertices, respectively, if C is a vertex bottleneck, then for any constant $\delta \in (0, 1/3]$, if $N_R > 4(2\times 5^{\frac{1-\delta}{2\delta}})N_C + 2$, there is a constant w_δ such that the routing time for piecewise time-independent Hamiltonians with minimum segment width $\Delta = w_\delta/\sqrt{N_L}$ is*

$$\mathrm{hrt}_\Delta(G) = \Omega\left(\frac{N_R^{1-\delta}}{\sqrt{N_L}N_C}\right). \tag{3.38}$$

Proof sketch: Our proof consists of two parts. In the first part, in Lemma 3.13, we show that any evolution $U(H,t)$ by an architecture-respecting piecewise time-independent Hamiltonian H with minimum segment width $w_\delta/\sqrt{N_L}$ for time $t \leq c_\delta N_R^{1-\delta}/(\sqrt{N_L}N_C)$ can be well-approximated by an architecture-respecting circuit $\widetilde{U}$ of depth $d \leq N_R/(4N_C) + 2 \times 5^{\frac{1-3\delta}{2\delta}}$ for some constant c_δ. To show this, we first prove Lemma 3.12, which bounds the approximation error of Trotter-Suzuki product-formula circuit approximations to $U(H,t)$ in systems with a tripartition. This lemma is based on the fact that vertex bottlenecks constrain the Frobenius norm of higher-order commutators of terms in any Hamiltonian that respects the tripartition connectivity, which we prove in Lemma B.1.1. We further make use of recent work [Zha+22] that bounds the error of Trotter-Suzuki product-formula circuit approximations in terms of the higher-order-commutator Frobenius norm. In the second part of the proof, we show that such a circuit cannot approximate a desired permutation unitary well. According to STE (Lemma 3.5), the amount of entanglement between L and $C \cup R$ that can be generated by an architecture-respecting circuit of depth d is bounded by $2dN_C$. There exist permutations p such that the permutation unitary U_p can be used to increase the entanglement between L and R by N_R, an increase which, by

STE, requires a circuit of depth $N_R/(2N_C)$. In Lemma 3.14, we show that such a circuit must be far from the target unitary U_p. Now combining the two parts of the proof, since we know that any architecture-respecting evolution $U(H,t)$ for $t \leq c_\delta N_R^{1-\delta}/\sqrt{N_L}N_C$ is well approximated by a circuit $\widetilde{U}$ that cannot perform the permutation p, our main result of a lower bound on the Hamiltonian routing time follows. We defer the formal proof to the end of this section.

In Appendix B.1, we show the following bound on the error of approximating architecture-respecting evolutions by circuits obtained using the Trotter-Suzuki formula.

Lemma 3.12. *Consider a graph with a vertex bottleneck. Let H be any time-independent Hamiltonian that respects this connectivity. Let $\widetilde{U}$ be the architecture-respecting simulation circuit corresponding to dividing the time-evolution $U = U(H,t)$ into M equal segments, and simulating each by the $(2k)^{\text{th}}$-order Trotter-Suzuki formula. Then, $\widetilde{U}$ has depth $d = 2 \times 5^{k-1}M$ [Zha+22] and there exists a function $g(k)$, only dependent on k, such that*

$$\|U - \widetilde{U}\|_{\text{F}} \leq g(k)\frac{t^{2k+1}}{M^{2k}}\sqrt{{N_L}^{2k}N_R N_C}. \tag{3.39}$$

We use this lemma to show that the circuit obtained from the Trotter-Suzuki formula can be an arbitrarily good approximation of the time-evolution.

Lemma 3.13. *Consider a graph with a vertex bottleneck. Let H be any piecewise time-independent Hamiltonian with minimum segment width Δ that respects the connectivity of a graph with a tripartition. Let $U = U(H,t)$ be the unitary corresponding to evolution by H for time t. Then, for any constant $\delta \in (0, 1/3]$ and $\epsilon > 0$, there exist constants $c_{\delta,\epsilon}, w_{\delta,\epsilon} > 0$ and*

$k \in \mathbb{N}_+$ *such that, for*

$$t \leq \frac{c_{\delta,\epsilon} N_R^{1-\delta}}{\sqrt{N_L} N_C} \tag{3.40}$$

and $\Delta > w_{\delta,\epsilon}/\sqrt{N_L}$*, there exists an architecture-respecting circuit* $\widetilde{U}$ *with depth* $d \leq N_R/(4N_C) + 2 \times 5^{k-1}$ *such that* $\|\widetilde{U} - U\|_F \leq \epsilon$.

Proof. We first assume H is time-independent for simplicity, and consider the piecewise case at the end of the proof. Let k be the smallest positive integer such that $\frac{1}{2k+1} \leq \delta$.

Selecting $M = \lceil \frac{1}{4} \left(2 \times 5^{k-1}\right)^{-1} N_R/N_C \rceil$, the circuit approximation $\widetilde{U}$ obtained from Lemma 3.12 has depth $d = 2 \times 5^{k-1} M \leq N_R/(4N_C) + 2 \times 5^{k-1}$. By Lemma 3.12,

$$\|U - \widetilde{U}\|_{\mathrm{F}} \leq 2^{6k} (5^{k-1})^{2k} g(k) \times t^{2k+1} \left(\frac{N_C}{N_R}\right)^{2k} \sqrt{{N_L}^{2k} N_R N_C}. \tag{3.41}$$

Bounding the right-hand side by ϵ and solving for t, we obtain

$$t \leq \frac{\left(2^{6k} 5^{2k^2-2k} g(k) \epsilon^{-1}\right)^{\frac{-1}{2k+1}} {N_R}^{1-\frac{1}{2k+1}}}{\sqrt{N_L} N_C}. \tag{3.42}$$

This proves the Lemma for time-independent H, because there is a suitable constant $c_{\delta,\epsilon}$ such that Eq. (3.40) implies Eq. (3.42), which guarantees $\|U - \widetilde{U}\|_{\mathrm{F}} \leq \epsilon$.

The above results easily generalize to piecewise time-independent Hamiltonians with minimum segment width $w_{\delta,\epsilon}/\sqrt{N_L}$. The Trotter error formula Eq. (3.39) does not hold directly because the Hamiltonian changes with time in some of the time windows of duration t/M. Nevertheless,

$$\frac{t}{M} \leq \frac{(8 \times 5^{k-1}) c_{\delta,\epsilon}}{N_R^{\delta} \sqrt{N_L}} < \frac{w_{\delta,\epsilon}}{\sqrt{N_L}} \tag{3.43}$$

for constant $w_{\delta,\epsilon}$ determined by $c_{\delta,\epsilon}$, which will be chosen shortly. Since t/M is smaller than the minimum segment width $w_{\delta,\epsilon}/\sqrt{N_L}$, each time window of duration t/M can be split into at most two smaller time windows, each containing time-independent evolution. Applying the Trotter error formula to these possibly smaller time windows, we obtain Eq. (3.39) again but with an extra prefactor of 2, because there are at most $2M$ new time windows of duration at most t/M. The extra factor of 2 propagates to Eq. (3.41), so we only need to choose a slightly smaller $c_{\delta,\epsilon}$ than the time-independent case (say, half of its value) to guarantee that Eq. (3.40) still implies $\|U - \widetilde{U}\|_{\mathrm{F}} \leq \epsilon$. Hence $w_{\delta,\epsilon} = (4 \times 5^{k-1}) c_{\delta,\epsilon}$ suffices. □

We now show that the approximation circuit $\widetilde{U}$ of Lemma 3.13 cannot achieve the permutation that routes the maximum number of qubits through the bottleneck.

Lemma 3.14. *For any graph G with a vertex bottleneck, consider a permutation p that maps m qubits from R to L through the vertex bottleneck and its associated permutation unitary U_p. Then, for any quantum circuit $\tilde{U}$ with depth $d < m/N_C$,*

$$\|\widetilde{U} - U_p\|_{\mathrm{F}}^2 > \frac{1}{4}\left(\frac{m - 2dN_C - 1}{N_R}\right)^2. \tag{3.44}$$

Proof. First, we bound $\|\widetilde{U} - U_p\|_{\mathrm{F}}^2$ in terms of the trace distance. For any bit string z of length N_R, we define

$$\rho_z := |z\rangle_R\langle z| \otimes (2^{-N_L - N_C} I)_{LC}. \tag{3.45}$$

One can purify the system by a $2^{N_L+N_C}$-dimensional ancilla (labeled by A) so that

$$\rho_z = \mathrm{tr}_A[|\Psi(z)\rangle\langle\Psi(z)|]. \tag{3.46}$$

Let $U' = U \otimes I_A$ and $\widetilde{U}' = \widetilde{U} \otimes I_A$. By definition of the Frobenius norm, and introducing $I = 2^{N_L+N_C} \sum_z \rho_z$,

$$\begin{aligned}
&\|U - \widetilde{U}\|_{\mathrm{F}}^2 \\
&= \frac{1}{2^{N_R}} \operatorname{tr}\left[\sum_z \rho_z |U - \widetilde{U}|^2\right] \\
&= \frac{1}{2^{N_R}} \sum_z \langle \Psi(z)| \, |U - \widetilde{U}|^2 \otimes I_A \, |\Psi(z)\rangle \\
&= \frac{1}{2^{N_R}} \sum_z \left[2 - 2\operatorname{Re}(\langle \Psi(z)| \, U'^{\dagger} \widetilde{U}' \, |\Psi(z)\rangle)\right]. \qquad (3.47)
\end{aligned}$$

Using the fact that $\operatorname{Re}(x) \leq |x|$, and for $x \in [0,1], 1 - x^2 \leq 2(1-x)$, we obtain

$$\begin{aligned}
&\sum_z \left[2 - 2\operatorname{Re}(\langle \Psi(z)| \, U'^{\dagger} \widetilde{U}' \, |\Psi(z)\rangle)\right] \\
&\geq \sum_z 2(1 - |\langle \Psi(z)| \, U'^{\dagger} \widetilde{U}' \, |\Psi(z)\rangle|) \\
&\geq \sum_z 1 - |\langle \Psi(z)| \, U'^{\dagger} \widetilde{U}' \, |\Psi(z)\rangle|^2 \\
&= \frac{1}{4} \sum_z \|U'|\Psi(z)\rangle\langle\Psi(z)|U'^{\dagger} - \widetilde{U}'|\Psi(z)\rangle\langle\Psi(z)|\widetilde{U}'^{\dagger}\|_1^2 \qquad (3.48) \\
&\geq \frac{1}{4} \sum_z \|\operatorname{tr}_A\left[U'|\Psi(z)\rangle\langle\Psi(z)|U'^{\dagger} - \widetilde{U}'|\Psi(z)\rangle\langle\Psi(z)|\widetilde{U}'^{\dagger}\right]\|_1^2 \\
&= \frac{1}{4} \sum_z \|\widetilde{U}\rho_z\widetilde{U}^{\dagger} - U_{\mathrm{p}}\rho_z U_{\mathrm{p}}^{\dagger}\|_1^2, \qquad (3.49)
\end{aligned}$$

where in Eq. (3.48), we have used the relation between fidelity and trace-norm for pure states.

By the small total entangling theorem (Lemma 3.5), $\widetilde{U}$ with depth d can increase the entan-

glement across the LC, R bipartition by at most $2dN_C$, i.e.,

$$S_R(\widetilde{U}\rho\widetilde{U}^\dagger) - S_R(\rho) \leq 2dN_C, \tag{3.50}$$

where S_R is the von Neumann entropy of the reduced density matrix on the right N_R qubits, and the vertex boundary of these qubits contains N_C qubits as in Fig. 3.2.

Observe that ρ_z is pure on R: $S_R(\rho_z) = 0$, so Eq. (3.50) implies

$$S_R(\widetilde{U}\rho_z\widetilde{U}^\dagger) \leq 2dN_C. \tag{3.51}$$

On the other hand, U_p routes m identity operators in LC to R, so that

$$S_R(U_\mathrm{p}\rho_z U_\mathrm{p}^\dagger) = m. \tag{3.52}$$

Intuitively, two unitaries that produce different changes in the von Neumann entropy of the same initial state must be distant. By the Fannes-Audenaert Inequality [Aud07] (the second inequality below),

$$\begin{aligned}
&\|\widetilde{U}\rho_z\widetilde{U}^\dagger - U_\mathrm{p}\rho_z U_\mathrm{p}^\dagger\|_1 \\
&\quad\geq \|\mathrm{tr}_{LC}[\widetilde{U}\rho_z\widetilde{U}^\dagger - U_\mathrm{p}\rho_z U_\mathrm{p}^\dagger]\|_1 \\
&\quad\geq \frac{S_R(U_\mathrm{p}\rho_z U_\mathrm{p}^\dagger) - S_R(\widetilde{U}\rho_z\widetilde{U}^\dagger) - 1}{\log \dim \mathcal{H}_R} \\
&\quad\geq \frac{m - 2dN_C - 1}{N_R}.
\end{aligned} \tag{3.53}$$

Using this to lower bound Eq. (3.49) gives the result. □

Combining Lemma 3.13 and Lemma 3.14, we may now prove our main result.

Proof of Theorem 3.11. Let p be a permutation that routes all N_R qubits from R to LC. Let U be the evolution by any architecture-respecting piecewise time-independent Hamiltonian H with minimum segment width $w_{\delta,\epsilon}/\sqrt{N}$ for time $t \leq c_{\delta,\epsilon} N_R^{1-\delta}/\sqrt{N_L} N_C$ for any $\delta > 0$. By Lemma 3.13, U can be ϵ-approximated by an architecture-respecting circuit $\widetilde{U}$ of depth at most $N_R/(4N_C) + 2 \times 5^{k-1}$ such that $\|\widetilde{U} - U\|_{\mathrm{F}} \leq \epsilon$. Now, Lemma 3.14 implies

$$\begin{aligned} \|\widetilde{U} - U_{\mathrm{p}}\|_{\mathrm{F}}^2 &\geq \frac{1}{4}\left(\frac{N_R - 2dN_C - 1}{N_R}\right)^2 \\ &\geq \frac{1}{4}\left(\frac{1}{2} - \frac{2(2 \times 5^{k-1})N_C}{N_R} - \frac{1}{N_R}\right)^2. \end{aligned} \tag{3.54}$$

We thus have

$$\begin{aligned} \|\widetilde{U} - U_{\mathrm{p}}\|_{\mathrm{F}} &\geq \frac{1}{4} - \frac{(2 \times 5^{k-1})N_C}{N_R} - \frac{1}{2N_R} \\ &\geq \frac{1}{4} - \frac{2(2 \times 5^{k-1})N_C + 1}{2N_R}. \end{aligned} \tag{3.55}$$

From the reverse triangle inequality, we have

$$\begin{aligned} \|U - U_{\mathrm{p}}\|_{\mathrm{F}} &\geq \left| \|\widetilde{U} - U_{\mathrm{p}}\|_{\mathrm{F}} - \|\widetilde{U} - U\|_{\mathrm{F}} \right| \\ &\geq \frac{1}{4} - \frac{2(2 \times 5^{k-1})N_C + 1}{2N_R} - \epsilon. \end{aligned} \tag{3.56}$$

Since k is the smallest positive integer such that $\frac{1}{2k+1} \leq \delta$,

$$\begin{aligned} &\frac{1}{2(k-1)+1} > \delta \\ &\implies k-1 < \frac{1-\delta}{2\delta}. \end{aligned} \tag{3.57}$$

Therefore,

$$\begin{aligned} N_R &> \frac{4(2\times 5^{\frac{1-\delta}{2\delta}})N_C + 2}{1-8\epsilon} \\ \implies N_R &> \frac{4(2\times 5^{k-1})N_C + 2}{1-8\epsilon} \\ \implies \|U - U_{\mathrm{p}}\|_{\mathrm{F}} &> \epsilon, \end{aligned} \tag{3.58}$$

and U cannot achieve the desired permutation. Selecting any constant ϵ allows us to pick $w_\delta = w_{\delta,\epsilon}$ for which the theorem holds. □

A straightforward application of Theorem 3.11 to the star graph yields the following result.

Corollary 3.15. *For any $\delta \in (0, 1/3]$, if $N > 8(2\times 5^{\frac{1-\delta}{2\delta}}) + 4$, the Hamiltonian routing time for the star graph on $N+1$ vertices with $\Delta = \frac{1}{\sqrt{N}}$ satisfies $\mathrm{hrt}_\Delta(S_{N+1}) = \Omega(N^{\frac{1}{2}-\delta})$.*

3.5 Entanglement flow through bottlenecks

In this section, we investigate the time taken to generate entanglement between subsystems connected by a bottleneck. In a system with a vertex bottleneck, the small total entangling theorem (Lemma 3.5) bounds the amount of entanglement that can be generated between L and RC

by a circuit of depth t as $\Delta S_L(\rho) \leq 2tN_C$. On the other hand, the small incremental entangling theorem (Lemma 3.4) [VMV13] and other results on entanglement flow [CVC05] bound the entangling rate in terms of the Hamiltonian operator norm as $\frac{\mathrm{d}S_L(\rho)}{\mathrm{d}t} \leq c\|H_{LC}\|N_C$. When $N_C \ll N_L$, these two bounds differ greatly. Integrating the bound from Lemma 3.4 for an evolution of time t, we obtain

$$\Delta(S_L(\rho)) \leq cN_LN_Ct. \tag{3.59}$$

Therefore, if this bound were saturable, by evolving for the same amount of time in a Hamiltonian-model rather than a gate-based model, we would obtain a speedup by a factor of N_L in our ability to generate entanglement.

At first, this seems very counter-intuitive. Indeed, in systems of free fermions, we showed a tighter bound of $O(N_C\sqrt{N_L})$ on the entangling rate between L and RC in Lemma 3.9. On the other hand, in systems of free bosons, the entangling rate can be very large, depending on the initial state. Consider a system with just two modes. If the initial state is the Fock state $|N, 0\rangle$, a simple calculation shows that applying a 50/50 beam splitter on the two modes leads to a mode entanglement of $\Theta(\log N)$.

A simple example illustrates that a large increase of the entanglement rate with continuous evolution may be possible for qubits as well. Consider the system depicted in Fig. 3.8. In a star graph, starting with an initial GHZ state on the leaves and $|0\rangle$ on the center, we can evolve into a GHZ state on the leaves and the center together in time π/N. Since the GHZ state has a von Neumann entropy of 1 between a single qubit and the remaining qubits, this implies that the entangling rate at some time must have been at least linear in N.

Although this system illustrates a worst-case counterexample for the instantaneous entan-

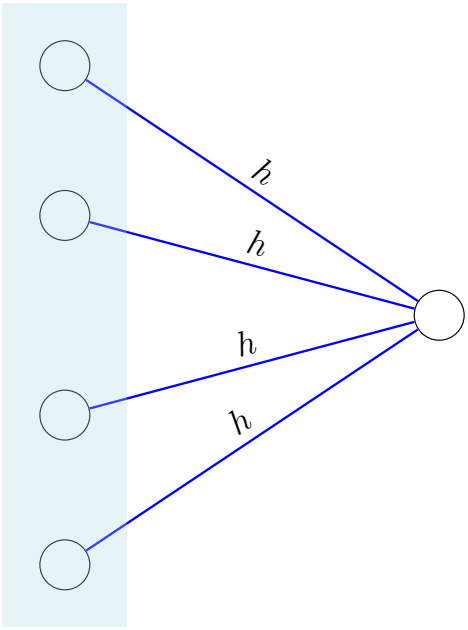


Figure 3.8: Linear entangling rate between N leaves of the star graph and the center vertex. If the initial state is a product state of the N-qubit GHZ state $\frac{1}{\sqrt{2}}(|00\ldots0\rangle + |11\ldots1\rangle)$ on the leaves, and $|0\rangle$ on the center, then evolving for time π/N with interaction $h = |1-\rangle\langle 1-|_{l_i,c}$ (i.e., the generator of a CNOT) on each edge results in a final $(N+1)$-qubit GHZ state on all the vertices.

gling rate through a bottleneck, the possibility of the bottleneck restricting our ability to produce entanglement between L and R remains open. Indeed, if it were always possible to generate fast entanglement, starting in any state, we could use this as a resource to perform routing. The existence of states that allow for fast entangling rates was used in Ref. [Bap+23] to obtain a fast routing protocol for the vertex barbell graph, although it also required very fast interactions with ancilla qubits. Our lower bound on the routing time in systems with a vertex bottleneck gives a hint that such states may be rare. Following this hint, we instead look at the entanglement that can be generated *on average* starting with a state drawn from a 1-design.

We define the entanglement capacity $C_\Delta(|\psi\rangle, t) := |S_L(U(H,t)|\psi\rangle\langle\psi|U(H,t)^\dagger - S_L(|\psi\rangle\langle\psi|)|$ as the amount of entanglement between L and RC that can be generated by an architecture-respecting piecewise time-dependent Hamiltonian H of minimum segment width Δ

in time t, starting in the state $|\psi\rangle$. Given a pure-state ensemble μ, we define the average entangling capacity as $\mathbb{E}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t)]$. We prove that, in a system with a vertex bottleneck, the average entangling capacity for any 1-design [AE07; GAE07; Ren+04] scales with N_L as $O(\sqrt{N_L})$. This can be contrasted with the worst-case scaling in N_L of $O(N_L)$ for the entanglement capacity that is obtained by integrating the entangling rate obtained from Lemma 3.4.

Theorem 3.16. *Let μ be a pure-state 1-design and let H be a piecewise time-dependent Hamiltonian with minimum segment width $\Delta = 1/\sqrt{N_L}$ that respects the connectivity constraints of a graph with a vertex bottleneck. Then for any $0 < \delta \leq 1/6$,*

$$\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t)] = O(1 + tN_C\sqrt{N_L}\,(N_L N_R/N_C)^\delta). \tag{3.60}$$

Proof. As in Section 3.4, let $U := U(H, t)$, and $\widetilde{U}$ be the circuit obtained by the $(2k)^{\text{th}}$ order Trotter-Suzuki formula for U, where k is chosen below. The entangling capacity of H on a particular input pure state $|\psi\rangle$ in time t is

$$\begin{aligned}
C_\Delta(|\psi\rangle, t) &= |S_L(U|\psi\rangle\langle\psi|U^\dagger) - S_L(|\psi\rangle\langle\psi|)| \\
&= \Big|S_L(U|\psi\rangle\langle\psi|U^\dagger) - S_L(\widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger) \\
&\qquad + S_L(\widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger) - S_L(|\psi\rangle\langle\psi|)\Big| \\
&\leq |S_L(U|\psi\rangle\langle\psi|U^\dagger) - S_L(\widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger)| + 2dN_C \\
&\leq 1 + N_L\|U|\psi\rangle\langle\psi|U^\dagger - \widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger\|_1 + 2dN_C,
\end{aligned} \tag{3.61}$$

where in the final line we have applied the Fannes-Audenaert inequality [Aud07].

Note that

$$\begin{aligned}
&\|U|\psi\rangle\langle\psi|U^\dagger - \widetilde{U}|\psi\rangle\langle\psi|\widetilde{U}^\dagger\|_1 \\
&= \|(U-\widetilde{U})|\psi\rangle\langle\psi|U^\dagger + \widetilde{U}|\psi\rangle\langle\psi|(U-\widetilde{U})^\dagger\|_1 \\
&\leq \|(U-\widetilde{U})|\psi\rangle\langle\psi|U^\dagger\|_1 + \|\widetilde{U}|\psi\rangle\langle\psi|(U-\widetilde{U})^\dagger\|_1 \qquad (3.62) \\
&= 2\sqrt{\langle\psi|\,(U-\widetilde{U})^\dagger(U-\widetilde{U})\,|\psi\rangle}, \qquad (3.63)
\end{aligned}$$

where in Eq. (3.62), we have used the triangle inequality, and Eq. (3.63) follows from the definition of the 1-norm. For states drawn from μ, applying Jensen's inequality and then using the fact that μ is a 1-design,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}\left[2\sqrt{\left\langle\psi\right|(U-\widetilde{U})^\dagger(U-\widetilde{U})\left|\psi\right\rangle}\right] \\
&\leq 2\sqrt{\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}\left[\langle\psi|\,(U-\widetilde{U})^\dagger(U-\widetilde{U})\,|\psi\rangle\right]} \\
&= 2\|U-\widetilde{U}\|_{\mathrm{F}}. \qquad (3.64)
\end{aligned}$$

Thus, we can bound the expected capacity by the Frobenius distance

$$\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle,t)] \leq 1 + 2dN_C + 2N_L\|U-\widetilde{U}\|_{\mathrm{F}}. \qquad (3.65)$$

As in Lemma 3.13 we first treat the case of time-independent H, and then show how the result generalizes to piecewise time-independent H. Using Lemma 3.12 with

$$M = \left\lceil t\sqrt{N_L}\,(N_L N_R/N_C)^{\frac{1}{4k+2}}\right\rceil, \qquad (3.66)$$

we obtain

$$\mathop{\mathbb{E}}_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t)] = O(1 + tN_C\sqrt{N_L}\,(N_L N_R/N_C)^{\frac{1}{4k+2}}) \tag{3.67}$$

since $d \leq 2 \times 5^{k-1}M$ [Zha+22]. The result for the time-independent case follows by choosing k large enough that $1/(4k+2) \leq \delta$.

Since $t/M \leq 1/\sqrt{N_L}$ is smaller than the minimum segment width, the Trotter error formula can be applied, and the result holds in the piecewise time-independent case with minimum segment width $1/\sqrt{N_L}$. We note that this restriction on the segment width arises for technical reasons (as in Eq. (3.43)). The number of segments M cannot be chosen to be very large, since this would allow for large d, giving a looser bound on the amount of entanglement generated. Our segment width is thus restricted such that t/M is smaller than the segment width without making M very large.

□

Though there exist examples of states with high entangling rate with a bottleneck (as in Fig. 3.8), Theorem 3.16 shows that in any 1-design μ there are few such states that can sustain a high entangling rate for long times. In particular, we can show that the entanglement rate for a randomly drawn state over any given period of time is unlikely to be significantly larger than $O(N_C\sqrt{N_L})$. Specifically, applying Markov's inequality and using Theorem 3.16, we obtain

$$\Pr_{|\psi\rangle\sim\mu}[C_\Delta(|\psi\rangle, t) \geq \Gamma t] = O\left(\frac{1 + tN_C\sqrt{N_L}\,(N_L N_R/N_C)^{\delta}}{\Gamma t}\right), \tag{3.68}$$

which is vanishingly small when

$$\Gamma \gg N_C \sqrt{N_L} \left(\frac{N_L N_R}{N_C} \right)^{\delta}. \tag{3.69}$$

For a state $|\psi\rangle$ drawn from μ, let $|\psi(t)\rangle = U(H,t)|\psi\rangle$ be a state along its trajectory under evolution by H. Eq. (3.68) shows that the averaged instantaneous entanglement rate over time,

$$C_\Delta(|\psi(0)\rangle, T) = \frac{1}{T} \int_0^T \frac{\mathrm{d}S_L(|\psi(t)\rangle\langle\psi(t)|)}{\mathrm{d}t} \mathrm{d}t, \tag{3.70}$$

is not greater than $O(N_C \sqrt{N_L}\, (N_L N_R / N_C)^{\delta})$ with high probability.

3.6 Discussion

In this chapter, we showed a lower bound that scales as $\Omega(N_R^{1-\delta}/\sqrt{N_L} N_C)$ for any $\delta > 0$ on the routing time in graphs with a vertex bottleneck. For the star graph, this provides a lower bound of $\Omega(\sqrt{N^{1-\delta}})$ on the routing time. We further showed an optimal routing protocol that saturates this bound in systems of free fermions on the star graph.

For graphs with a vertex bottleneck, the best previous lower bound we are aware of is $\Omega(1)$, following from the Small Incremental Theorem [VMV13] or a Lieb-Robinson bound [CLY23]. However, unlike those results, a limitation of our work is our introduction of additional assumptions about the allowed Hamiltonian: piecewise time-independence, absence of ancillas, and bounds on the on-site terms $\|h_i\|$. A straightforward argument can be used to prove a similar lower bound for routing with time-dependent Hamiltonians for which the Frobenius norm has suitably bounded derivative, by simulating the time-dependent Hamiltonian evolution with

a piecewise time-independent evolution, as in [Ber+14]. A natural direction for future work is to extend our lower bound to systems with more general Hamiltonians with more general time dependence or unbounded local terms. Our assumption on the piecewise time dependence of the allowed Hamiltonians stems from our use of a Trotterized circuit with Frobenius commutator norm scaling of error in Lemma 3.13. If bounds on the Trotter error for random input states can be extended beyond piecewise time-independent Hamiltonians to time-dependent Hamiltonians whose norm has unbounded derivative, our result could be extended to arbitrary time-dependent Hamiltonian routing as well. Such an approach might also extend our result to Hamiltonians with unbounded local (on-site) terms, since we could move into the interaction picture, removing the on-site terms and adding time-dependence to the non-local terms. We conjecture that a similar lower bound can be proved for more general time-dependent Hamiltonians (whose norms may have large derivatives), and that our assumed bound on the on-site terms can be removed. Another direction for future work is to investigate routing with a piecewise time-independent Hamiltonian of smaller segment width. For technical reasons, our result is restricted to segment width $\propto 1/\sqrt{N_L}$. It is natural to ask whether routing with a piecewise time-independent Hamiltonian of arbitrarily small segment width is equivalent to routing with arbitrary time-dependent Hamiltonians (i.e., whether $\lim_{\Delta\to 0} \mathrm{hrt}_\Delta(G) = \mathrm{hrt}(G)$).

We also assumed that interactions between any two sites in the same partition have bounded norm. This is a natural assumption in systems where interactions between the partitions are physically similar to interactions within the partitions. A possible direction for future work would be to relax this assumption. We expect that this would lead to a looser lower bound on the routing time. Indeed, in systems with ancillas and unbounded interactions within each partition, we can perform routing faster. An example of this is the W-state broadcasting-based protocol of Bapat

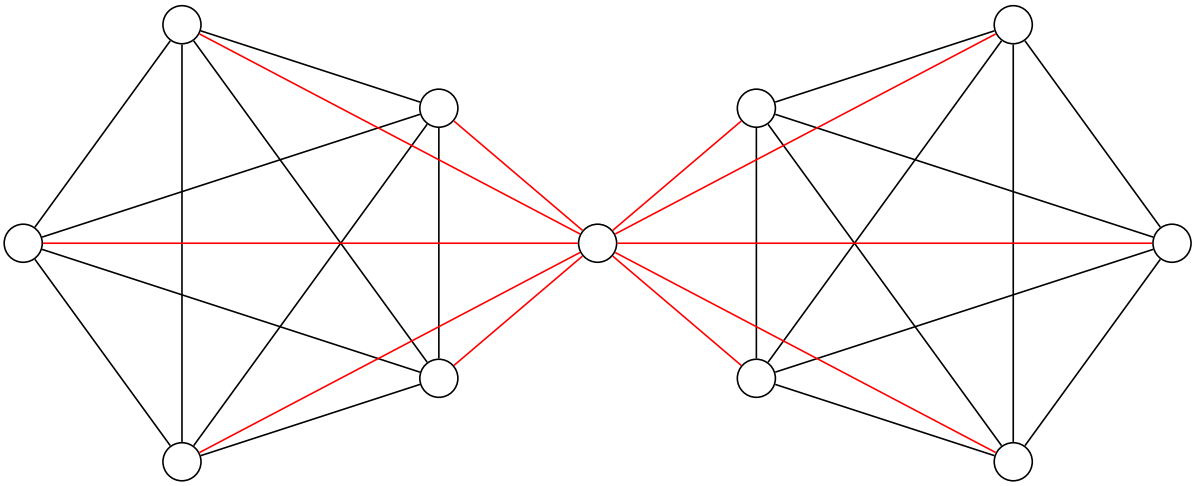


Figure 3.9: The vertex barbell graph with $2N + 1$ vertices consists of two copies of the complete graph K_N connected by a common vertex. Here, we depict the vertex barbell with 11 vertices.

et al. [Bap+23], which performs Hamiltonian routing in time $O(\sqrt{N})$ on the vertex barbell graph with $2N + 1$ vertices (Fig. 3.9), obtaining a speedup over the gate-based model (where routing requires depth $\Theta(N)$). Recent work [Yin25] implies the possibility of even faster routing, in $\mathrm{poly}(\log N)$ time, in the same system. However, speedups obtained in this routing model may not always apply since they require the experimentally challenging capability of fast swap operations between qubits and their ancillas. Our techniques provide an $\Omega(N^{1/2-\delta})$ lower bound in the more realistic setting of the vertex barbell graph without ancillas. Even with bounded 2-site interactions, a related question for future work is to obtain lower bounds that apply when ancillas are permitted.

A more refined analysis might be used to remove the δ dependence of our qubit lower bound. While our proofs hold for arbitrarily small positive constant δ, we conjecture that the result holds with $\delta = 0$. A limitation of our techniques is that our lower bound has a constant factor inversely proportional to δ, so taking the limit as $\delta \to 0$ causes the lower bound to diverge.

We also showed an upper bound on the average entanglement capacity, or average amount

of entanglement that can be produced in time t, through a vertex bottleneck. Intuitively, one expects a bottleneck to limit entangling rates, but surprisingly there are states for which fast entanglement with a central qubit is possible. One such example is the W state, which can be used to create constant entanglement with a central vertex in time $O(1/\sqrt{N})$ [Bap+23]. The GHZ state allows for an even higher entangling rate, producing a constant amount of entanglement in time $O(1/N)$. Our bound on the average entangling capacity for a 1-design shows that states with an entangling rate higher than $\Omega(\sqrt{N})$ are rare, and constrains their use for tasks such as routing. This result may be interpreted as showing that most states cannot support fast entangling through a vertex bottleneck. An open question is whether this bound can be extended to mixed states as well, which are known to exhibit surprising entanglement dynamics [Cub+03]. It would also be interesting to investigate how this bound constrains other quantum information processing tasks based on their requirements of entanglement. Similarly to our routing lower bound, our entanglement capacity upper bound can be extended to time-dependent Hamiltonians whose norm has bounded derivative by approximating their evolution by a piecewise time-dependent one. Natural questions for future work are to extend the entanglement capacity upper bound to Hamiltonians with arbitrary time dependence, with fast on-site interactions and fast interactions within partitions, and with ancillas. We also conjecture that this bound holds with $\delta = 0$.

The existence of states that saturate the worst-case bounds on entangling rate, which depend on the operator norm, necessitates our proof by the average case, using the Frobenius norm. For quantum information tasks involving operators with significant difference in the operator norm and Frobenius norm (as for nested commutators of Hamiltonians with a vertex bottleneck), our results provide an example where studying the average case yields tighter bounds than the worst case.

Recent work by one of the authors has shown a fast approximate broadcasting (i.e., GHZ encoding) protocol on the complete graph [Yin25]. While this protocol has not been proven to have asymptotically vanishing error, numerics suggest this may be the case. Such a protocol, coupled with fast ancilla interactions, would allow for approximate routing on the vertex barbell in time $\mathrm{poly}(\log N)$. An open question is to resolve whether such broadcasting can be done with vanishing error, or even exactly. The best known exact protocol for routing on the vertex barbell (with fast local ancillas) involves broadcasting into a W state [Bap+23], and takes time $O(\sqrt{N})$.

While we developed optimal protocols for routing free particles, a natural open question is whether such protocols can be sped up by allowing interaction terms or even terms that do not preserve particle number, such as $l_i^\dagger c_i^\dagger$. In bosonic systems, introduction of a certain type of interaction term has been shown to provide speedups for the closely related task of state transfer [VKS24].

In Section 3.3.2, we discussed the difficulty of extending our free-particle routing protocols to work for qubits or qudits. A major open question is to obtain fast (sublinear-time) qubit routing protocols through bottlenecks that leverage continuous time evolution. While such results have been obtained on the barbell graph using W states (and we expect that the result of Yin [Yin25] allows for even faster routing using GHZ states), these protocols require the use of ancillas with fast interactions. Currently, we do not know of any sublinear-time routing protocol for qubits on the star graph.

Chapter 4: The complexity of thermalization in finite quantum systems

4.1 Introduction

Statistical mechanics has been remarkably successful at describing the physics of large numbers of particles in thermal equilibrium. Experimental outcomes of systems in thermal equilibrium can be predicted using statistical ensembles based on macroscopic observables, independently of the system's microstate. Therefore, the process through which an isolated system evolves toward thermal equilibrium, known as thermalization, has been a central topic of investigation in condensed matter physics. In classical systems, the emergence of thermalization has been attributed to ergodicity and chaotic dynamics [Deu18]. Ergodic systems evolve such that they eventually come arbitrarily close to every point in phase space on a constant energy surface. As classical measurements can be modeled as long-time averages, the long-time behavior of ergodic systems is hence similar to that of a uniform distribution over the constant-energy submanifold of phase space, also known as the classical microcanonical ensemble [PB21; Rue99]. The true relationship between thermalization and ergodicity in classical systems is subtle and nuanced. Ergodicity is difficult to prove and has only been rigorously shown for some systems [Sin70; Bun79; Sim04; CM06], notably for systems with suitably strong chaotic behavior [Ott02; Ozo89]. Indeed, it has been argued that ergodicity is neither necessary nor sufficient for thermalization [Bri95]. Nevertheless, ergodicity has remained a useful conceptual tool to understand

thermalization in classical systems.

However, the ergodic explanation of thermalization can't be naturally extended to isolated quantum systems. In particular, unitary time evolution does not sample all wavefunctions within a given energy range of the starting state, and hence cannot be ergodic. As an alternative, the Eigenstate Thermalization Hypothesis (ETH) [Sre94; Deu91; Deu18] has been proposed to explain the exhibition of thermalization in isolated quantum systems. ETH proposes that observables with slowly varying diagonal elements (in the energy eigenbasis) and small off-diagonal elements approach their microcanonical thermal values. However, the ETH has not been proven to be generally applicable. Indeed, systems exist that do not thermalize, due to phenomena such as many-body localization [Aba+19], quantum many-body scars [Tur+18], and integrability [Deu18; KWW06] (although one can consider integrable systems as equilibrating to generalized Gibbs ensembles that take their additional conservation laws into account [Lan+15; VR16]).

A large volume of theoretical and experimental work has gone into determining which properties and initial states ensure thermalization [Tas16; RDO08; KIH14; GE16; Cam+25]. A natural question thus arises: How can we determine whether a given quantum system thermalizes? While this question has been studied by physicists, it was first considered from the viewpoint of computational complexity theory by Shiraishi and Matsumoto [SM21], who proved that even for product initial states with translationally invariant Hamiltonians, the problem of determining whether a system thermalizes or not is undecidable in the limit of an infinitely large system. Their results utilize techniques inspired by previous work in the field of Hamiltonian complexity proving complexity and computability results in physics [CPW15; Moo90].

However, in the lab — and nature more generally — we only have access to physical systems of finite size. This raises the question of how difficult it is to predict thermalization of these

finite-sized systems. While the result of Shiraishi and Matsumoto [SM21] proving undecidability in the thermodynamic limit, along with the more general hardness of simulating quantum systems, suggests this may be a difficult task, it is entirely possible that there exists an efficient algorithm for it. The result of Shiraishi and Matsumoto [SM21] makes use of a reduction from an undecidable problem (the Halting problem) to the problem of determining whether a system thermalizes, but undecidable instances may require a large (unbounded) size and therefore may not be seen in systems of finite size such as those encountered in nature and the lab.

In this work we consider the problem of determining whether a *finite-sized* quantum system relaxes to a given value, or thermalizes. To study this from a complexity-theoretic perspective, we formulate relaxation as a decision problem.

Definition 4.1 ((Informal) Finite-Size Relaxation, FSRelax)**.** Consider an input of a k-local Hamiltonian H acting on N qudits with constant local Hilbert space dimension d, a sum of local observables $\mathcal{A}_N = \frac{1}{N}\sum_i A_i$, an input classical description of product state $|\psi_0\rangle$, a value A^*, and constants c, ϵ. Let the long time average of $\mathcal{A}_N$ be

$$\bar{\mathcal{A}}_N \coloneqq \lim_{T\to\infty} \frac{1}{T}\int_0^T dt \langle\psi_0|e^{iHt}\mathcal{A}_N e^{-iHt}|\psi_0\rangle \tag{4.1}$$

Then, output: YES if $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$ or NO if $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$.

FSRelax is depicted in Fig. 4.1.

One relevant state of interest that describes the thermal behavior of an isolated system is the microcanonical ensemble, which represents the set of possible states within a fixed range of energy of our initial state.

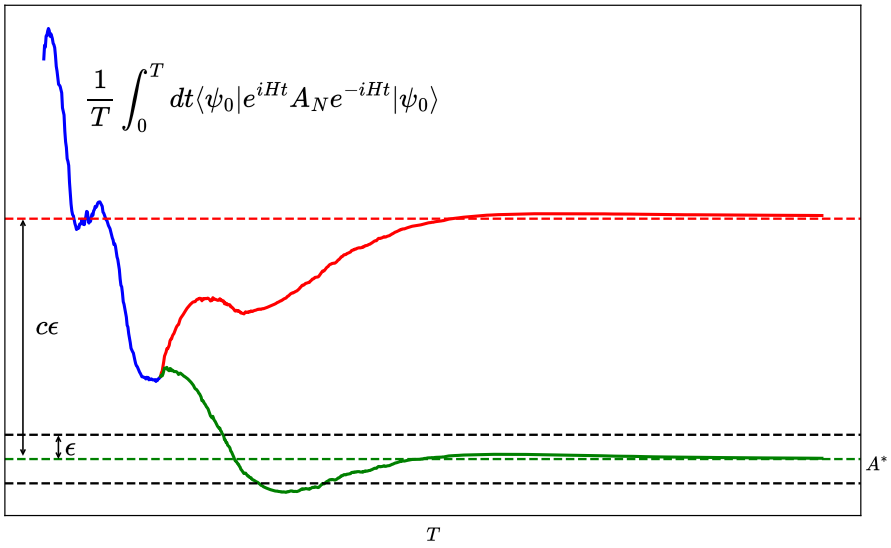


Figure 4.1: The relaxation decision problem. For a given initial state $|\psi\rangle$, we plot the time-average of $\mathcal{A}_N \coloneqq \frac{1}{T}\int_0^T dt\langle\psi_0|e^{iHt}\mathcal{A}_N e^{-iHt}|\psi_0\rangle$. A^* is indicated by the dashed green line. Given the promise that $\bar{\mathcal{A}}_N \coloneqq \lim_{T\to\infty}\langle\mathcal{A}_N\rangle_T$ is such that either $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$ or $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$, FSRelax is the problem of deciding which of these two holds.

Definition 4.2 (Microcanonical Ensemble)**.** For a Hamiltonian H with eigenstates $\{|\lambda_i\rangle\}$ and corresponding eigenenergies $\{\lambda_i\}$ the quantum microcanonical ensemble $\rho_{\mathsf{MC}}(H, E)$ at energy E is defined as

$$\rho_{\mathsf{MC}}(H, E) = \frac{1}{W}\sum_i f\left(\frac{\lambda_i - E}{w}\right)|\psi_i\rangle\langle\psi_i|, \tag{4.2}$$

where f is an energy window function (typically decaying rapidly in its argument), $W = \sum_i f\left(\frac{\lambda_i - E}{w}\right)$ and w is a characteristic energy width.

To investigate thermalization, we define the following decision problem:

Definition 4.3 ((Informal) Finite-Size Thermalization to the microcanonical state, FSTherm(MC))**.** Consider an input of a k-local Hamiltonian H acting on N qudits with local Hilbert space dimension d, a sum of local observables $\mathcal{A}_N = \frac{1}{N}\sum_i A_i$, an input classical

description of product state $|\psi_0\rangle$, and constants c, ϵ. Let the long time average of $\mathcal{A}_N$ be

$$\bar{\mathcal{A}}_N \coloneqq \lim_{T\to\infty} \frac{1}{T} \int_0^T dt \langle\psi_0|e^{iHt}\mathcal{A}_N e^{-iHt}|\psi_0\rangle. \tag{4.3}$$

Then, output: YES if $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{MC}}(E)]| \leq \epsilon$ or NO if $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{MC}}(E)]| \geq c\epsilon$, where E is the expected energy of the input state $\langle\psi_0|H|\psi_0\rangle$.

We additionally consider a closely related problem to $\mathsf{FSTherm(MC)}$ which is the problem of thermalization to the Gibbs state $(\frac{e^{-\beta H}}{\mathrm{tr}[e^{-\beta H}]})$.

Definition 4.4 ((Informal) Finite-Size Thermalization to the Gibbs state, $\mathsf{FSTherm(Gibbs)}$)**.** Consider an input of a k-local Hamiltonian H acting on N qudits with local Hilbert space dimension d, a sum of local observables $\mathcal{A}_N = \frac{1}{N}\sum_i A_i$, an input classical description of product state $|\psi_0\rangle$, and constants c, ϵ. Let the long time average of $\mathcal{A}_N$ be

$$\bar{\mathcal{A}}_N \coloneqq \lim_{T\to\infty} \frac{1}{T} \int_0^T dt \langle\psi_0|e^{iHt}\mathcal{A}_N e^{-iHt}|\psi_0\rangle. \tag{4.4}$$

Then, output: YES if $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{G}}]| \leq \epsilon$ or NO if $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{G}}]| \geq c\epsilon$, where where ρ_{G} is the Gibbs state of H with expected energy $E \coloneqq \langle\psi_0|H|\psi_0\rangle$.

Our main results are that $\mathsf{FSRelax}, \mathsf{FSTherm(MC)}$ are computationally intractable, even for a quantum computer and for simple systems. We first prove that the problem of determining whether a given observable relaxes to a particular value is PSPACE-complete.

Theorem 4.5. $\mathsf{FSRelax}$ *is PSPACE-complete for a translationally invariant, nearest neighbour Hamiltonian with* $d \geq 51$*, with a 1-local translationally invariant observable* A_i*.*

We then show a PSPACE algorithm to compute the microcanonical expectation value of a local observable, which allows us to make use of the PSPACE algorithm to determine FSRelax to decide FSTherm(MC) in PSPACE. We also show that the FSTherm(MC) problem is contained in PSPACE. By allowing a slightly larger local Hilbert space dimension, one can choose A_i such that $\bar{\mathcal{A}}_N = \mathrm{tr}[\rho_{\mathsf{MC}}\mathcal{A}_N]$, where ρ_{MC} is the microcanonical ensemble of H. Therefore, we have:

Theorem 4.6. *There exists a translationally invariant, nearest neighbour Hamiltonian and 1-local, translationally invariant observable such for which* $A^* = \mathrm{tr}[\mathcal{A}_N\rho_{MC}]$*, such that solving* FSTherm(MC) *is* PSPACE*-hard under quantum polynomial time reductions. In addition, for any such Hamiltonian and observable,* FSTherm(MC) *is contained in* PSPACE.

We note that our proof of PSPACE-hardness under quantum polynomial time reductions has the technical limitation that we restrict to a class of microcanonical states with width $w = \Omega(\|H\|/\sqrt{\log N})$. Further details on this restriction and the possibility of removing it in future work are discussed in Section 4.6.2 and Section 4.8.

In the case of thermalization to the Gibbs state, we use similar techniques to show that FSTherm(Gibbs) is contained in PSPACE.

Theorem 4.7. FSTherm(Gibbs) $\in$ PSPACE.

We further present evidence and a physically justified conjecture that FSTherm(Gibbs) is PSPACE-hard. Assuming our conjecture, we may obtain PSPACE-hardness for FSTherm(Gibbs) under *classical* polynomial time reductions. Hence, if this conjecture (Conjecture 4.43) is true, FSTherm(Gibbs) is PSPACE-complete.

Finally, we show that the experimentally relevant problem of determining whether finite-sized systems of qubits thermalize in polynomial time is BQP-complete.

Our results show evidence for the difficulty of understanding the mechanisms of relaxation and thermalization, and it's intractability from a complexity theoretic perspective even for systems in nature or the lab. Theorem 4.5 demonstrates that we should not expect there to be an efficient way of finding (or even approximating) the time-averaged expectation value of local observables. Furthermore, Theorem 4.6 shows that in general any "easy-to-check" condition which ensures thermalization will not apply to all Hamiltonians. The systems we design may be of further physical interest, since our results show that even simple systems (translationally invariant, 1D nearest-neighbour Hamiltonians and initial product states) can exhibit highly complex relaxation and thermalization dynamics.

The rest of this chapter is organized as follows: In Section 4.2 we define relevant complexity classes and introduce the relevant mathematical quantities for thermalization. In Section 4.3 we prove the PSPACE-hardness of FSRelax. In Section 4.4 we prove that FSRelax $\in$ PSPACE. Combined with our result on PSPACE-hardness, this shows that FSRelax is PSPACE-complete. In Section 4.5, we prove thatFSTherm(MC) $\in$ PSPACE and FSTherm(Gibbs) $\in$ PSPACE. In Section 4.6, we derive the quantum polynomial time reduction from FSTherm(MC) to FSRelax,. Together, these prove our main result Theorem 4.6. We also discuss a conjecture which, if true, would imply that FSTherm(Gibbs) is PSPACE-complete. Finally, in Section 4.7, we prove that deciding whether a finite size system relaxes in finite time is BQP-complete. We conclude with a discussion of our results and outlook for future work in Section 4.8.

4.2 Preliminaries

We briefly introduce classical and quantum Turing machines. A more in-depth treatment can be found in Arora and Barak [AB09] and Watrous [Wat99]. Turing machines consist of a tape with cells onto which symbols can be written, as well as a tape head, which both stores a state and points to a specific tape cell. The tape head is able to read the symbol on the cell, and based on the read symbol and the head's state, (potentially) both write to the cell and move the head to an adjacent cell, according to a transition function. A quantum Turing machine can be defined similarly, but with a quantum tape and the ability to perform quantum operations (Toffoli gates, Hadamard gates, phase-shift gates, or single-qubit measurements in the computational basis) on the quantum tape [Wat08]. A promise problem is a pair (A_{yes}, A_{no}) where $A_{yes}, A_{no} \subseteq \{0,1\}*$ are sets of strings satisfying $A_{yes} \cap A_{no} = \varnothing$ [Wat08]. We are primarily interested in the following classes of promise problems. The first class PSPACE, is the class of decision problems solvable by a deterministic Turing Machine in polynomial space. More formally:

Definition 4.8 (PSPACE)**.** A promise problem $A = (A_{yes}, A_{no})$ is in PSPACE (polynomial space) if and only if there exists a deterministic Turing machine running in polynomial space that accepts every string $x \in A_{yes}$ and rejects every string $x \in A_{no}$.

PSPACE contains the commonly known classes P, BQP, NP, and it is believed that PSPACE is not contained in them (and thus PSPACE-complete problems are believed to be harder to solve). There is a corresponding class for quantum computers, BQPSPACE, which is roughly the class of decision problems solvable by a quantum Turing Machine in polynomial space.

Definition 4.9 (BQPSPACE,[Wat08])**.** A promise problem $A = (A_{yes}, A_{no})$ is in BQPSPACE

(bounded-error quantum polynomial space) if and only if there exists a quantum Turing machine running in polynomial space that accepts every string $x \in A_{yes}$ with probability at least 2/3 and accepts every string $x \in A_{no}$ with probability at most 1/3.

It is known that BQPSPACE = PSPACE [Wat08].

Definition 4.10 (PrQSPACE [Wat08])**.** A promise problem $A = (A_{yes}, A_{no})$ is in PrQSPACE (probabilistic quantum polynomial space) if and only if there exists a quantum Turing machine M running in polynomial space that accepts every string $x \in A_{yes}$ with probability strictly greater than $1/2$ and accepts every string $x \in A_{no}$ with probability at most $1/2$.

It is known that PrQSPACE = PSPACE [Wat08].

4.2.1 Algorithmic primitives

We make use of two algorithmic primitives: block encodings and quantum phase estimation. Block encodings allow the implementation of arbitrary sub-normalized matrices as the upper left block of a unitary [Gil+19; CKS17; Ber+15; CGJ19; LC19]. We restate Definition 43 from [Gil+19] here:

Definition 4.11 (Block Encoding)**.** Suppose that A is an s-qubit operator, $\alpha, \epsilon \in \mathbb{R}_{>0}$, and $a \in \mathbb{N}$. Then an $(s+a)$-qubit unitary U is an (α, a, ϵ)-block encoding of A if and only if

$$\|A - \alpha((I \otimes \langle 0^a|)U(I \otimes |0^a\rangle))\| \leq \epsilon \tag{4.5}$$

Let s represent the s-qubit register, and a represent the ancillas used by the block-encoding.

Then, for an initial state $|\psi\rangle$,

$$U|\psi\rangle = A'_s|\psi\rangle|0^a\rangle_a + |\phi^\perp\rangle \tag{4.6}$$

where $\langle 0^a|_a|\phi^\perp\rangle = 0$ and $\|A - A'_s\| \leq \epsilon$. Usually, when we make use of a block-encoding, we will postselect on the ancillas being in the state $|0^a\rangle$, since this heralds the correct application of A'_s to $|\psi\rangle$.

We now restate some standard results on the complexity of preparing block encodings of (functions of) matrices.

Lemma 4.12 (Block Encoding Sparse Matrices (Lemma 48 of [Gil+19]))**.** *Let $A \in M_{2^w}(\mathbb{C})$ be a s-sparse matrix with $|A_{ij}| \leq 1$ for all i, j, and for which both the position of nonzero entries and these non-zero entries' values have efficient classical descriptions (i.e., can be computed in polynomial time classically). Then, there exists a quantum circuit implementing a $(s, w+3, \varepsilon)$ block encoding of A with $O\left(\operatorname{poly} w + \operatorname{poly}\log \frac{s^2}{\varepsilon}\right)$ fundamental gates and $w + O\left(\operatorname{poly}\log \frac{s^2}{\varepsilon}\right)$ space.*

Lemma 4.13 (Polynomial Functions of Block Encodings (Thm. 56 of [Gil+19]))**.** *Let $U \in U(2^n)$ be a (α, a, ε)-block encoding of some Hermitian A with a circuit that uses $\mathscr{A}_U$ ancillas, $\mathscr{C}_U$ fundamental gates, and whose circuit has a classical description computable in time $\mathscr{T}_U$. Further, let $p \in \mathbb{R}[x]$ be a polynomial of degree d such that $\sup_{x\in[-1,1]} |p(x)| \leq 1/2$. Then, for any $\delta \geq 0$ there exists a circuit using $O(d(a + \mathscr{C}_U))$ fundamental gates and $O(d\mathscr{A}_U)$ ancillas to produce a $(1, a+2, 4d\sqrt{\varepsilon/\alpha} + \delta)$ encoding $\tilde{U}_\delta$ of $p(A/\alpha)$. A classical description of this circuit can be computed in time $O(d\mathscr{T}_U + \operatorname{poly} d + \operatorname{poly}\log \delta^{-1})$.*

To make full use of this, we need some polynomial approximations of specific functions we wish to block encode.

Lemma 4.14 (Polynomial Approximation of Square Root)**.** *For all $\epsilon \leq 1/2$, there exists an efficiently computable polynomial $P^{(\epsilon)}_{sqrt}(x) \in \mathbb{C}[x]$ of degree $O(\log \epsilon^{-1})$ such that*

$$\sup_{x\in[-1,1]} \left| P^{(\epsilon)}_{\text{sqrt}}(x) - \sqrt{1+x/2} \right| \leq \epsilon. \tag{4.7}$$

Proof. Note that for $x \in [-1,1]$, $\sqrt{1+\frac{x}{2}} = \sum_{n=0}^{\infty} \frac{1}{1-2n}\binom{2n}{n}\left(-\frac{x}{8}\right)^n$. We have that $\sum_{n=0}^{\infty} \frac{1}{|1-2n|}\binom{2n}{n}2^{-3n} = 2 - 2^{-1/2} \approx 1.293$.

Thus, by [Gil+19, Cor. 66], we have that for any $0 < \epsilon \leq 1/2 < 1/(2-2^{-1/2})$, there exists a polynomial $P^{(\epsilon)}_{\text{sqrt}}(x)$ of degree $O(\log \epsilon^{-1})$ (in the language of the cited corollary, $B = 2-2^{-1/2}$, $\delta = r = 1/2$, and $x_0 = 0$) that satisfies $\sup_{x\in[-1,1]} \left| \sqrt{1+x/2} - P^{(\epsilon)}_{\text{sqrt}}(x) \right| \leq \epsilon$. □

Lemma 4.15 (Polynomial Approximations of Gaussian)**.** *For any $0 < \epsilon \leq 1$, there exists an efficiently computable polynomial $P^{(\epsilon)}_{Gauss(w)}(x) \in \mathbb{C}[x]$ of degree $O(\max\{w^{-1}, \sqrt{\log \epsilon^{-1}}\} \times \sqrt{\log \epsilon^{-1}})$ such that $\sup_{x\in[-1,1]} |P^{(\epsilon)}_{\text{Gauss}(w)} - e^{-\frac{\pi x^2}{2w^2}}| \leq \epsilon$.*

Proof. Note that if $x \in [-1,1]$, $\frac{\pi}{2}\left(\frac{x}{w}\right)^2 \in [0, \frac{\pi}{2w^2}]$. Further, note that by [SV14, Thm. 4.1], there exists a polynomial $r_{\epsilon,w}$ of degree $O\left(\sqrt{\max\left\{\frac{\pi}{2w^2}, \log \epsilon^{-1}\right\} \log \epsilon^{-1}}\right)$ such that $\sup_{x\in[0,\frac{\pi}{2w^2}]} |e^{-x} - r_{\epsilon,w}(x)| \leq \epsilon$. Thus, letting $P^{(\epsilon)}_{\text{Gauss}(w)}(x) := r_{\epsilon,w}\left(\frac{\pi x^2}{w^2}\right)$, we arrive at our result. □

Next, we state a result about the number of ancillas necessary to implement phase estimation to m bits of accuracy with total error $\leq \epsilon$ [KSV02; Cle+98].

Lemma 4.16 (Space Complexity of Phase Estimation)**.** *Given an n-qubit state $|\psi\rangle$, an n-qubit unitary $U \in U(2^n)$ such that $U|\psi\rangle = e^{2\pi i\phi}$, $m \in \mathbb{N}$, and $\epsilon \in (0,1)$ there exists a quantum circuit taking as input $|\psi\rangle|0^{m+1+\lceil 2\log_2 \epsilon^{-1}\rceil}\rangle$ and using no additional ancillae which outputs a state $|\psi\rangle|\tilde{\phi}^{(m,\epsilon)}\rangle$ such that $\|\Pi_{\phi,m}|\tilde{\phi}^{(m,\epsilon)}\rangle\| \leq \epsilon$, where $\Pi_{\phi,m} := \sum_{\substack{\mathbf{b}\in\{0,1\}^{m+1+\lceil 2\log_2\epsilon^{-1}\rceil} \\ |0.b_1b_2\cdots-\phi|>2^{-(m+1)}}} |\boldsymbol{b}\rangle\langle\boldsymbol{b}|$.*

In addition, we will need the following corollary about the approximate orthogonality of estimates of distant eigenvalues.

Corollary 4.17. *Let ϕ_1, ϕ_2 be eigenphases of U as above, such that $|\phi_1 - \phi_2| > 2^{-m}$ for some $m \in \mathbb{N}$. Then, if $|\phi_1^{(m,\epsilon)}\rangle, |\phi_2^{(m,\epsilon)}\rangle$ are as defined above,*

$$\left|\left\langle \phi_1^{(m,\epsilon)} \middle| \phi_2^{(m,\epsilon)} \right\rangle\right| \leq 2\epsilon. \tag{4.8}$$

4.2.2 Thermalization

Throughout, we will restrict ourselves to local Hamiltonians of the form:

$$H = \sum_{i=1}^{m} h_i \tag{4.9}$$

where $m = \text{poly } N$, h_i acts on at most an $O(1)$ number of qudits, and $\|h_i\| = O(1)$.

Let A_i be a local (single-qudit) observable acting on qudit i, then define:

$$\mathcal{A}_N = \frac{1}{N} \sum_{i=1}^{N} A_i. \tag{4.10}$$

We will further mainly consider translationally invariant $\bar{\mathcal{A}}_N$, such that each $A_i = A$. Given an initial state $|\psi_0\rangle$ and a Hamiltonian H, the long time-average of $\mathcal{A}_N$ is:

$$\bar{\mathcal{A}}_N \coloneqq \lim_{T\to\infty} \frac{1}{T} \int_0^T dt \langle\psi_0|e^{iHt}\mathcal{A}_N e^{-iHt}|\psi_0\rangle \tag{4.11}$$

The problem we will be interested in is, roughly, given an initial state $|\psi_0\rangle$, what expectation

value of $\mathcal{A}_N$ does this state relax to after an infinite amount of time? In order to study this from a complexity theoretic viewpoint, we define this as a decision problem.

Definition 4.18 ((Formal) Finite Size Relaxation FSRelax)**.** Consider a k-local Hamiltonian H acting on N qudits with constant local Hilbert space dimension d, a local observable A, and a fixed value A^*. Let c, ϵ be positive constants. Let the long time average of A be as given in Eq. (4.11).

Input: A classical description of product state $|\psi_0\rangle$.

Output: YES if $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$ or NO if $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$.

Promise: Either $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$ or $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$. Let $\{|\lambda_i\rangle\}$ be the eigenstates of H. Then, the initial state $|\psi_0\rangle := \sum_i c_i |\lambda_i\rangle$ is only supported on energy eigenstates with an inverse exponential gap, i.e. $\exists\, m = O(\text{poly}(N))\; s.t. \forall\, i, j$ if $c_i, c_j \neq 0\,, |\lambda_i - \lambda_j| \geq d^{-poly(m)}$.

FSRelax decides whether the observable $\mathcal{A}_N$ for the input initial state relaxes over a long time to the target value A^*. We note that the promise on the initial state strictly makes the problem easier, and so the problem without the promise is at least as hard.

The problem of determining whether the system thermalizes is closely related to determining relaxation, where the observable relaxes specifically to the expected value for a thermodynamic ensemble. We consider two relevant ensembles: the microcanonical ensemble (Definition 4.2) and the Gibbs ensemble. The expectation value of a local observable O with respect to the microcanonical ensemble is:

$$\langle O \rangle_{\text{MC}} := \text{tr}[O \rho_{\text{MC}}(H, E)] = \frac{1}{W} \sum_i f\left(\frac{\lambda_i - E}{w}\right) \langle \lambda_i | O | \lambda_i \rangle \tag{4.12}$$

where $W = \sum_i f\left(\frac{\lambda_i - E}{w}\right)$ is a normalizing factor and $\{|\lambda_i\rangle\}$ are the eigenstates of the relevant

Hamiltonian.

In the rest of the chapter, we select $f(x) = e^{-\pi x^2}$, such that

$$\rho_{\mathsf{MC}}(H, E) = \frac{1}{\sum_i e^{-\pi\left(\frac{\lambda_i - E}{w}\right)^2}} \sum_i e^{-\pi\left(\frac{\lambda_i - E}{w}\right)^2} |\lambda_i\rangle\langle\lambda_i| \tag{4.13}$$

By choosing a Gaussian window function, we ensure that the distribution of energies in the microcanonical ensemble is peaked at E and decays exponentially with distance. While in principle, the microcanonical state could be defined with a rectangular window function, this has the disadvantage of making the state sensitive to the precision of w in the case where there are eigenenergies close to the boundaries $|E \pm w|$. Previous work [AG19] has used projectors onto the eigenspace above or below a certain value to prepare the Gibbs states, but with the requirement that the value is sufficiently far from the spectrum of H. In order to avoid the requirement of a promise that the boundaries $|E \pm w|$ are away from the spectrum of H and to avoid non-analyticities, we make use of a smooth window function. For ease of notation, and since we are usually only concerned with the microcanonical ensemble at an energy corresponding to that of a single initial state, in the rest of the chapter we denote $\rho_{\mathsf{MC}}(E)$ as ρ_{MC}.

Another state which can represent a thermalized system is the canonical ensemble, or Gibbs state, which represents the state at a fixed temperature $(1/\beta)$.

Definition 4.19 (Canonical Ensemble)**.** For a Hamiltonian H the canonical ensemble, or Gibbs state ρ_{G} at inverse temperature β is defined as

$$\rho_{\mathsf{G}} = \frac{e^{-\beta H}}{\mathrm{tr}[e^{-\beta H}]}, \tag{4.14}$$

When we consider thermalization to the Gibbs state, β is implicitly determined by the expected energy of the initial state $|\psi_0\rangle$, such that $\frac{\mathrm{tr}[He^{-\beta H}]}{\mathrm{tr}[e^{-\beta H}]} = E$, where $E = \langle\psi_0|H|\psi_0\rangle$. Since $\frac{\mathrm{tr}[He^{-\beta H}]}{\mathrm{tr}[e^{-\beta H}]}$ is monotonic in β, there exists a unique β and hence unique ρ_{G} satisfying this constraint.

We now define the decision problems for Thermalization:

Definition 4.20 ((Formal) Finite Size Thermalization to the microcanonical ensemble $\mathsf{FSTherm(MC)}$)**.** Consider a k-local Hamiltonian H acting on N qudits with local Hilbert space of constant dimension d, an observable $\mathcal{A}_N$ which is a sum of local observables. Let c, ϵ be constants. Let the long time average of $\mathcal{A}_N$ be as given in Eq. (4.11).

Input: A classical description of product state $|\psi_0\rangle$.

Output: YES if $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{MC}}(E)]| \leq \epsilon$ or NO if $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{MC}}(E)]| \geq c\epsilon$, where E is the expected energy of the input state $\langle\psi_0|H|\psi_0\rangle$.

Promise: Either $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{MC}}(E)]| \leq \epsilon$ or $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{MC}}(E)]| \geq c\epsilon$. Let $\{|\lambda_i\rangle\}$ be the eigenstates of H. Then, the initial state $|\psi_0\rangle := \sum_i c_i|\lambda_i\rangle$ is only supported on energy eigenstates with an inverse exponential gap, i.e. $\exists\, m = O(\mathrm{poly}(N))\, s.t.\ \ \forall\, i \neq j$ if $c_i, c_j \neq 0\,, |\lambda_i - \lambda_j| \geq d^{-poly(m)}$.

Definition 4.21 ((Formal) Finite Size Thermalization to the Gibbs ensemble $\mathsf{FSTherm(Gibbs)}$)**.** Consider a k-local Hamiltonian H acting on N qudits with local Hilbert space of constant dimension d, an observable $\mathcal{A}_N$ which is a sum of local observables. Let c, ϵ be constants. Let the long time average of $\mathcal{A}_N$ be as given in Eq. (4.11).

Input: A classical description of product state $|\psi_0\rangle$.

Output: YES if $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{G}}]| \leq \epsilon$ or NO if $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{G}}]| \geq c\epsilon$, where ρ_{G} is the Gibbs state of H with expected energy equal to that of the input state $\langle\psi_0|H|\psi_0\rangle$.

Promise: Either $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{G}}]| \leq \epsilon$ or $|\bar{\mathcal{A}}_N - \mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{G}}]| \geq c\epsilon$. Let $\{|\lambda_i\rangle\}$ be the eigenstates of H. Then, the initial state $|\psi_0\rangle := \sum_i c_i |\lambda_i\rangle$ is only supported on energy eigenstates with an inverse exponential gap, i.e. $\exists\, m = O(\mathrm{poly}(N))\; s.t.\;\; \forall\, i \neq j$ if $c_i, c_j \neq 0\,, |\lambda_i - \lambda_j| \geq d^{-poly(m)}$.

We will also be interested in a slightly different version of the relaxation problem, where we promise that the Hamiltonian and initial state relax in polynomial amount of time in the system size. The study of this "quickly" relaxing set of Hamiltonians is motivated by the fact that these are the set of Hamiltonians for which we can realistically study relaxation and equilibriation in a lab. We call this version "Finite-Time, Finite-Space Relaxation", FTFSRelax:

Definition 4.22 (Finite-Time, Finite-Space Relaxation, FTFSRelax)**.** Consider a local Hamiltonian H acting on N qudits with local Hilbert space dimension d, a local observable A, a fixed value A^* and a timescale $T = O(\mathrm{poly}\, N)$. Let c, ϵ be constants. Let the long time average of A be as given in Eq. (4.11).

Input: A classical description of product state $|\psi_0\rangle$.

Output: YES if $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$ or NO if $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$.

Promise: Either $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$ or $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$. Let $\{|\lambda_i\rangle\}$ be the eigenstates of H. Then, the initial state $|\psi_0\rangle := \sum_i c_i |\lambda_i\rangle$ is only supported on energy eigenstates with an inverse exponential gap, i.e. $\exists\, m = O(\mathrm{poly}(N))$ such that $\forall\, i, j$ if $c_i, c_j \neq 0$ then $|\lambda_i - \lambda_j| = \Omega(1/\,\mathrm{poly}\, N)$.

4.3 PSPACE-hardness of FSRelax

We show the PSPACE-hardness of FSRelax by reducing a computationally hard problem, the finite-size halting problem FSHalt, to FSRelax.

Definition 4.23 (Finite Size Halting (FSHalt))**.** The problem FSHalt takes as input a Turing machine M, an input x and an integer n in unary. It outputs YES if M halts on x within n tape space, and NO if it does not.

By definition, PSPACE consists of all languages for which membership can be decided by a deterministic Turing machine in polynomial space. For any language $L \in$ PSPACE decided by a Turing machine M we can define a TM M' that runs M, and halts on a Yes instance, or runs infinitely on a No instance. The language L can thus be decided by FSHalt with M', the instance x, and $n = \text{poly}\,|x|$ as input. Further, FSHalt $\in$ PSPACE since we can simply simulate M with a universal Turing machine restricted to polynomial space. Therefore, FSHalt is PSPACE-complete.

We present a brief overview of our proof strategy to show that FSRelax is PSPACE-hard, which is closely related to the construction used by Shiraishi and Matsumoto [SM21]. Our reduction relies on constructing a Hamiltonian that encodes the action of a Universal Reversible Turing Machine (URTM), and the input state to FSRelax encodes the input to the URTM. We design our Hamiltonian and observable such that the long-time average of the observable depends on whether the URTM halts or not. Therefore, the relaxation value of the observable allows us to determine the solution to instances of FSHalt, and thus FSRelax must be as hard as FSHalt. Later, in Section 4.6, we show how to modify this construction to apply to FSTherm(MC). We

discuss the relationship between our hardness construction and that of Shiraishi and Matsumoto [SM21] in Section 4.8.

4.3.1 Turing Machines and Setup

We construct a Turing machine $\mathcal{M}$ composed of three TMs:

- TM1, a 10-state, 8-symbol URTM [Mor17]. The states of the finite control are Q_u and the symbols for the tape are Γ_u
- TM2, which has 1 state r and 2 symbols (a_1, a_2)
- TM3, which has 10-states Q_{rev}, and uses the same 8-symbols Γ_u to run TM1 in reverse.

Each site of our physical system represents one cell or the finite control of the TM tape. Our system is chosen to have periodic boundary conditions, so our tape is also periodic. Similar to Shiraishi and Matsumoto [SM21], we denote the cells with symbols from TM1 and TM3 as M-cells, and the cells with symbols of TM2 as A-cells.

Our TM $\mathcal{M}$ functions as follows. Assume in the initial state we fix a constant α as the fraction of cells which are to be M-cells, and we evenly distribute M and A cells. The initial state has the input to TM1 directly on the M cells. Every A cell starts with the symbol a_1. TM1 commences on the M cells, and when it encounters an A cell it simply skips over in the same direction. TM1 functions as a URTM on the M cells. If TM1 halts, the finite control state changes to the state r and TM2 commences, flipping every A cell from a_1 to a_2. Eventually, TM2 wraps around the tape and encounters an A cell with the symbol a_2 – this is the first cell encountered by TM2, adjacent to the halting cell of TM1. Then, the finite control state is changed to a state

of TM3, which starts on the halting cell of TM1 and runs TM1 in reverse, before finally halting. By running TM1 in reverse, but with the A cells flipped, we ensure that in the case of halting, the A cells remain flipped for a sufficiently long duration. TM3 has a special halting state which it enters at the point when it would transition to the initial state of TM1.

In order to ensure that our Hamiltonian is 2-local, we split each step of TM1 and TM3 into 2 - first to change the control state, and then to move the control either left or right. Thus the set of symbols and states required is:

$$\begin{aligned} Q &= \{W, M\} \times (Q_u \cup Q_{rev}) \cup \{r\}, \\ \Gamma &= \Gamma_u \cup \{a_1, a_2\}, \end{aligned} \tag{4.15}$$

the "W" corresponding to the write substep and the "M" corresponding to the move substep.

We note $|Q_u| = 10, |\Gamma_u| = 8$, and $|Q_{rev}| = 10$. To ensure the combined TM functions as intended, we furnish the combined TM with an additional transition rule which takes it out of the halting state of TM1, and moves its control to the state r, thereby initiating TM2, as well as a transition rule for when TM2 encounters a_2, which then halts TM2 and initiates TM3. Thus, if TM1 does not halt on its input, $\mathcal{M}$ has all A-cells filled with a_1, and if TM1 halts on its input, $\mathcal{M}$ has all A-cells filled with a_2.

4.3.2 TM-to-Hamiltonian Mapping

We now need find a way of mapping our Turing Machine $\mathcal{M}$ to a Hamiltonian. Fortunately, this is a well studied problem, originally in Ref. [Fey86], and a variety of TM-to-Hamiltonian mappings have been developed for proving hardness of properties of Hamiltonians

[BT14; BCO17; Wat19].

Let U be the unitary which implements the transition rule of a given TM. In particular, if there is a transition rule $|ab\rangle \rightarrow |cd\rangle$, then we define the term $U_{abcd} = |cd\rangle\langle ab|$, applied to all relevant qudits. We then define the set of transition rules to be $\mathcal{T}$ and the corresponding Hamiltonian to be:

$$H = \sum_{abcd \in \mathcal{T}} (U^{\dagger}_{abcd} + U_{abcd}). \tag{4.16}$$

Now let $|w_0\rangle$ be an initial state of some computation in the computational basis. We can then define the state after the k^{th} application of the TM transition rule as $|w_k\rangle = V^k|w_0\rangle$, where $V := \sum_{abcd \in \mathcal{T}} U_{abcd}$ maps product states representing some Turing machine configuration to those of the Turing machine's incrementation by the reversibility of the transition relations described by $\mathcal{T}$.

Given a particular input state $|w_0\rangle$, the encoded computation either (a) halts, (b) loops forever. We can apply a unitary transformation to this get the Hamiltonian in a block-diagonal form, where each of the blocks corresponds to a different input state (see [KSV02] or [Wat19, Section 5]). Within each block, the Hamiltonian has the structure of a graph Laplacian of either a line graph (halting case) or loop (non-halting case). Thus for a given input state, we can define an effective Hamiltonian H_{eff} corresponding to the particular input state:

$$H_{\text{eff}} = \sum_{k=0}^{T-2} |w_k + 1\rangle\langle w_k| + h.c., \tag{4.17}$$

where T is either the halting time, or the time until the TM begins to loop. This effective Hamil-

tonian has eigenstates of the adjacency matrix of a line graph or loop which are [YC08; DK20]:

$$\lambda_j(H_{\text{eff}}) = 2\cos\left(\frac{(j+1)\pi}{T+1}\right) \tag{4.18}$$

for for halting and

$$\lambda_j(H_{\text{eff}}) = 2\cos\left(\frac{2\pi j}{T}\right) \tag{4.19}$$

for non-halting, $j \in \{0, 1, \ldots, T-1\}$. Due to the determinism of Turing Machine evolution and the finite number of configurations of the machine, we see that $T = O(d^N)$.

4.3.3 Long-Time Expectation Values

Given the TM-to-Hamiltonian mapping presented in Section 4.3.2, we can encode the TMs given in Section 4.3.1. We now consider an initial state:

$$|w_0\rangle = |q_0\rangle \otimes |y\rangle, \tag{4.20}$$

where q_0 corresponds to the the TM head in its initial state and $y \in \Gamma^N$, with a proportion $(1-\alpha)$ evenly spaced out A-cells in the state $|a_1\rangle$. On the sites that are not set to a_1, y will encode Turing machine we wish to determine FSHalt for into the URTM's tape symbols. We then choose the TM1 to run the problem FSHalt . If the TM1 halts, we have it transition to TM2 which flips $|a_1\rangle \to |a_2\rangle$. For the remainder of Section 4.3.1, we will define $A_i := |a_2\rangle\langle a_2|_i$.

Lemma 4.24. *If y encodes a Turing Machine that either halts in time T_h or never halts, we have*

that

$$\bar{\mathcal{A}}_N = \begin{cases} \frac{1-\alpha}{2}\left[1 - \frac{2}{2T_h+N+1}\right] & \textit{halting} \\ 0 & \textit{non-halting}. \end{cases} \tag{4.21}$$

Proof. We begin by considering the limit definition for $\bar{\mathcal{A}}_N$:

$$\begin{aligned} \bar{\mathcal{A}}_N &= \lim_{T\to\infty} \frac{1}{T}\int_0^T dt \langle w_0|e^{iHt}\mathcal{A}_N e^{-iHt}|w_0\rangle \\ &= \lim_{T\to\infty} \sum_{ij} \frac{1}{T}\int_0^T dt e^{i(\lambda_i-\lambda_j)t} c_i^* c_j \langle\lambda_i|\mathcal{A}_N|\lambda_j\rangle \\ &= \sum_{ij} \delta_{\lambda_i,\lambda_j} c_i^* c_j \langle\lambda_i|\mathcal{A}_N|\lambda_j\rangle \\ &= \sum_{ij} \delta_{i,j} c_i^* c_j \langle\lambda_i|\mathcal{A}_N|\lambda_j\rangle \\ &= \sum_{j} |c_j|^2 \langle\lambda_j|\mathcal{A}_N|\lambda_j\rangle, \end{aligned} \tag{4.22}$$

where for the penultimate equality we have used that none of the eigenvalues of the effective Hamiltonian have the same energy (as per Eq. (4.18)).

If TM1 never halts, then TM2 never starts and none of then a_1 cells are ever flipped to a_2 and we see that

$$\mathcal{A}_N|w_j\rangle = 0, \quad \forall j. \tag{4.23}$$

This in turn means $\mathcal{A}_N|\lambda_j\rangle = 0$ for all j, as all $|\lambda_j\rangle$ are superpositions of the $|w_k\rangle$. Thus, for non-halting, $\bar{\mathcal{A}}_N = 0$. Shiraishi and Matsumoto [SM21, Eq. S.46] show that for halting, $\bar{\mathcal{A}}_N$

becomes:

$$\sum_j |c_j|^2 \langle \lambda_j | \mathcal{A}_N | \lambda_j \rangle = \frac{3}{2(T+1)} \left(\langle w_0 | \mathcal{A}_N | w_0 \rangle + \langle w_{T-1} | \mathcal{A}_N | w_{T-1} \rangle \right) + \frac{1}{T+1} \sum_{j=1}^{T-2} \langle w_j | \mathcal{A}_N | w_j \rangle - \frac{1}{2(T+1)} \sum_{\substack{0 \leq j,j' \leq T-1 \\ j'=j\pm 2}} \langle w_j | \mathcal{A}_N | w_{j'} \rangle. \tag{4.24}$$

We now analyze the terms of this equation. We have that the M cells are evenly spaced. We assume that every M cell is followed by $\alpha^{-1} - 1$ A-cells, and that αN of these α^{-1} blocks constitute the whole Turing Machine. Thus, noting that $\langle w_0 | \mathcal{A}_N | w_0 \rangle = 0$, $\langle w_{T-1} | \mathcal{A}_N | w_{T-1} \rangle = 1 - \alpha$, and the final term in Eq. (4.24) is 0, it remains to calculate

$$\begin{aligned} \sum_{j=1}^{T-2} \langle w_j | \mathcal{A}_N | w_j \rangle &= (1-\alpha) \left[T - 2 - (T_h + N + 1) + 1 \right] + \sum_{j=T_h+1}^{T_h+N} \langle w_j | \mathcal{A}_N | w_j \rangle \\ &= (1-\alpha) \left[T - 2 - (T_h + N) \right] + \sum_{j=T_h+1}^{T_h+N} \langle w_j | \mathcal{A}_N | w_j \rangle. \end{aligned} \tag{4.25}$$

If $|w_j\rangle$ has l A-cells with the symbol a_2, then

$$\langle w_j | \mathcal{A}_N | w_j \rangle = l \langle a_2 | A | a_2 \rangle. \tag{4.26}$$

Due to the uniformity of the placement of the M-cells (and recalling our assumption that $\alpha N \in$

$\mathbb{Z}$), for $j \in [T_h + 1, T_h + N]$,

$$\begin{aligned}
\sum_{j=T_h+1}^{T_h+N} \langle w_j|\mathcal{A}_N|w_j\rangle &= \sum_{j=1}^{N} \frac{[j - \lfloor \alpha j \rfloor]}{N} \langle a_2|A|a_2\rangle \\
&= \frac{1}{N}\left[\sum_{j=1}^{N} j - \sum_{j=1}^{N} \lfloor \alpha j \rfloor\right] \langle a_2|A|a_2\rangle \\
&= \frac{1}{N}\left[\frac{N(N+1)}{2} - \sum_{j=0}^{N-1} \lfloor \alpha j \rfloor - N\right] \langle a_2|A|a_2\rangle \\
&= \frac{1}{N}\left[\frac{N(N+1)}{2} - \alpha^{-1} \sum_{k=0}^{\alpha N-1} k - N\right] \langle a_2|A|a_2\rangle \\
&= \left[\frac{N+1}{2} - \frac{1}{\alpha N}\frac{\alpha N(\alpha N-1)}{2} - 1\langle a_2|\right] A|a_2\rangle \\
&= \frac{(1-\alpha)N}{2} \langle a_2|A|a_2\rangle. \qquad (4.27)
\end{aligned}$$

Thus,

$$\sum_{j=1}^{T-2} \langle w_j|\mathcal{A}_N|w_j\rangle = (1-\alpha)[T - (T_h + N/2 + 2)]\langle a_2|A|a_2\rangle. \qquad (4.28)$$

Combining this with Eq. (4.24) and Eq. (4.25), we obtain

$$\bar{\mathcal{A}}_N = (1-\alpha)\frac{T - (T_h + N/2 + 1/2)}{T+1} \langle a_2|A|a_2\rangle. \qquad (4.29)$$

Finally, noting that $T = 2T_h + N$, and $A = |a_2\rangle\langle a_2|$ we obtain

$$\begin{aligned}
\bar{\mathcal{A}}_N &= (1-\alpha)\frac{T_h + N/2 - 1/2}{2T_h + N + 1} \\
&= \frac{1-\alpha}{2}\left[1 - \frac{2}{2T_h + N + 1}\right]. \qquad (4.30)
\end{aligned}$$

□

4.3.4 Putting it All Together and Verifying the Promise

We now wish to use our previously proved results about the Hamiltonian to prove hardness of FSRelax. For the remainder of this section, we remember that A satisfies $\langle a_2|A|a_2\rangle = 1$.

Theorem 4.25. FSRelax *is PSPACE-hard for* $d = 51$.

Proof. We consider the Hamiltonian generated from the TM-to-Hamiltonian mapping, where the TM is chosen to be that from Section 4.3.1. We then choose an initial state $|\psi_0\rangle = |q_0\rangle \otimes |y\rangle$, where y is a string $y \in \Gamma^{\times N-1}$. In particular, we choose y to have a fraction α M cells, where we are free to choose α. We choose the M first n cells to be a bit string of length n, $x \in \{0,1\}^n$.

We then choose the first TM encoded to run the FSHalt problem. The action of the final TM only occur if the initial TM halts. We see from Lemma 4.24 that the time-averaged expectation value is:

$$\bar{\mathcal{A}}_N = \begin{cases} \frac{1-\alpha}{2}\left[1 - \frac{2}{2T_h+N+1}\right] & \text{halting} \\ 0 & \text{non-halting.} \end{cases} \tag{4.31}$$

By choosing $\alpha = \Theta(1)$ to be a sufficiently small constant we get, for sufficiently large N, separated values of $\bar{\mathcal{A}}_N$ in the halting and non-halting cases respectively. Since FSHalt is PSPACE-hard, then determining which case will occur for $\bar{\mathcal{A}}_N$ is also PSPACE-hard. We show later in Lemma 4.26 that the promise that $|\lambda_i - \lambda_j| > d^{-\operatorname{poly} m}$ is satisfied by this Hamiltonian and input state. □

To find the local Hilbert space dimension of the Hamiltonian, we simply need to add up the

necessary Hilbert space size for the TMs to function. We realise that the total Hilbert space size will be the size of all the alphabet symbols plus the size of all the TM head state symbols. Thus we expect the local Hilbert space dimension to be:

$$\begin{aligned} d &= 2 \times (|Q_u| + |Q_{rev}|) + |\{r\}| + |\Gamma_u| + |\{a_1, a_2\}| \\ &= 51. \end{aligned} \tag{4.32}$$

Finally we verify that the Hamiltonian and initial state we have used for our hardness construction satisfies the promise of FSRelax.

Lemma 4.26 (Promise is Satisfied)**.** *Let $|w_0\rangle$ be an initial basis state, and let H be a Hamiltonian from a TM-to-Hamiltonian mapping on N qudits. Then:*

$$\min_{ij} |\lambda_i(H_{\mathit{eff}}) - \lambda_j(H_{\mathit{eff}})| = \Omega\left(d^{-2N}\right). \tag{4.33}$$

Proof. From Eq. (4.18), we see that the minimum difference in energies occurs when for $j = i + 1$, giving:

$$\begin{aligned} |\lambda_{j+1}(H_{\text{eff}}) - \lambda_j(H_{\text{eff}})| &= 2\left(\cos\left(\frac{(j+1)\pi}{T+1}\right) - \cos\left(\frac{j\pi}{T+1}\right)\right) \\ &= 4\left|\sin\frac{\pi}{2(T+1)}\sin\frac{(j+1)\pi}{2(T+1)}\right| \\ &\geq \frac{\pi^2}{4(T+1)^2} \\ &= \Omega(T^{-2}) = \Omega(d^{-2N}). \end{aligned} \tag{4.34}$$

□

4.4 Containment of FSRelax in PSPACE

We show that FSRelax $\in$ PSPACE by demonstrating a $\operatorname{poly} N$-space quantum algorithm to estimate $\bar{\mathcal{A}}_N$. Combined with the result that BQPSPACE $\subseteq$ PSPACE [Wat99], and using the fact that FSTherm(MC) is an instance of FSRelax, this shows that FSTherm(MC) $\in$ PSPACE.

Lemma 4.27. *Let $\mathcal{A}_N$ be defined as in Eq. (4.10). For any $\epsilon \in [0, 1/4]$, there exists a quantum circuit to prepare a $(4, N\lceil \log_2 d \rceil + 5, \epsilon)$-block encoding of $\sqrt{1 - \mathcal{A}_N/2dN}$ using $O(\operatorname{poly} \log N \operatorname{poly} \log \epsilon^{-1})$ ancillas, with a classical description that can be computed in time $O(\operatorname{poly} N \operatorname{poly} \log \epsilon^{-1})$.*

Proof. We begin by noting that $\mathcal{A}_N$ is a dN-sparse matrix, and that all its entries are less than or equal to 1 in magnitude. Thus, by Lemma 4.12, we have that we can produce a $(dN, \lceil \log_2(d) \rceil N + 3, \epsilon_2)$-block encoding of $\mathcal{A}_N$ with $O\Big(\operatorname{poly} * \log\left(\frac{d^2N^2}{\epsilon_2}\right)\Big)$ ancillas.

By Lemma 4.14, for all $\epsilon_1 \leq 1/4$, there exists a polynomial $p(x)$ of degree $O(\log \epsilon_1^{-1})$ such that

$$\sup_{x \in [-1,1]} \left| \frac{1}{4}\sqrt{1 - x/2} - p(x) \right| \leq \epsilon_1, \tag{4.35}$$

and $\sup_{x\in[-1,1]} |p(x)| \leq 1/2$. Thus, by Lemma 4.32, we can prepare a $\Big(1, \lceil \log_2(d) \rceil N + 5, O\Big(\log \epsilon_1^{-1} \sqrt{\epsilon_2/dN} + \delta\Big)\Big)$-block encoding of $p\left(\frac{\mathcal{A}_N}{dN}\right)$ with $O\Big(\log \epsilon_1^{-1} \operatorname{poly} * \log\left(\frac{d^2N^2}{\epsilon_2}\right)\Big)$ ancillas, $O(\log \epsilon_1^{-1}(\operatorname{poly} N + \operatorname{poly} \log \epsilon_2^{-1}))$ fundamental gates, and a classical description that can be computed in time

$$O\big(\log \epsilon_1^{-1} \left(\operatorname{poly} N + \operatorname{poly} * \log \epsilon_2^{-1}\right) + \operatorname{poly} * \log \epsilon_1^{-1} + \operatorname{poly} \log \delta^{-1}\big). \tag{4.36}$$

Thus, to obtain the desired result, it suffices to take $\delta, \epsilon_1, \log \epsilon_1^{-1} \sqrt{\epsilon_2 / dN} = \Theta(\epsilon)$. To do this, we see that we can take $\delta, \epsilon_1 = \Theta(\epsilon)$, and $\epsilon_2 = \Theta\left(\left(\frac{\sqrt{N}\epsilon}{\log \epsilon^{-1}}\right)^2\right)$.

□

We use this block encoding circuit to estimate $\bar{\mathcal{A}}_N$.

Lemma 4.28. *For any $0 < \epsilon \leq 1/4$, there is a quantum algorithm that takes space $O(\epsilon^{-2} \operatorname{poly} \log \epsilon^{-1} \operatorname{poly} N)$ to return Γ such that*

$$|\Gamma - \bar{\mathcal{A}}_N| \leq \epsilon \tag{4.37}$$

with probability of at least $2/3$.

Proof. Note that we can uniquely write our initial state as $|\psi_0\rangle = \sum_i c_i |\lambda_i\rangle$ with $|\lambda_i\rangle$ a λ_i-eigenvector of H such that $\lambda_i \neq \lambda_j$ for all $i \neq j$. Further, by the promise condition, we have that there exists $m = \operatorname{poly} N$ that satisfies $\min\left\{d^{-N}, \inf_{i \neq j} |\lambda_i - \lambda_j|\right\} > 2^{-m}$ for all i for which $c_i \neq 0$. We begin by noting that we can implement a unitary $\tilde{U}_{\epsilon_1}$ such that $\|\tilde{U}_{\epsilon_1} - e^{iH}\| \leq \epsilon_1 2^{-5(m+1)}$ in polynomial space with an efficient classical description via standard Trotterization techniques [Chi+21]. If we perform phase estimation on $\tilde{U}_{\epsilon_1}$ with $\mathcal{O}(m)$ bits of phase precision, we would then get an approximation to the phase estimation circuit on e^{iH} with (operator norm) error at most ϵ_1.

Let $|\tilde{\lambda}_i^{(m, \epsilon_2 2^{-(1+m)})}\rangle$ be the output in the phase register of running quantum phase estimation of e^{iH} on $|\lambda_i\rangle$, desired to be accurate to m-bits of precision with error parameter $\epsilon_2 2^{-(m+1)}$. By Lemma 4.16, we have that this can be achieved by a circuit with $3(m+1) + \lceil 2 \log_2 \epsilon_2^{-1} \rceil$ total qubits. Despite the required exponential depth of the phase estimation circuit, it has an efficient

classical description due to its repeated structure - so a description of each layer of gates can be computed in polynomial time classically. Then, by Corollary 4.17 we have:

$$\left|\langle\tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})}|\tilde{\lambda}_j^{(m,\epsilon_2 2^{-(m+1)})}\rangle - \delta_{i,j}\right| \leq \epsilon_2 2^{-m}, \tag{4.38}$$

and hence,

$$\left|\sum_{i,j=0}^{|\sigma(H)|-1} c_i c_j^* \langle\lambda_j|\mathcal{A}_N|\lambda_i\rangle\langle\tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})}|\tilde{\lambda}_j^{(m,\epsilon_2 2^{-(m+1)})}\rangle - \sum_{i=0}^{|\sigma(H)|-1} |c_i|^2\langle\lambda_i|\mathcal{A}_N|\lambda_i\rangle\right| \leq \epsilon_2. \tag{4.39}$$

The long time average of $\mathcal{A}_N$ for our initial state $|\psi_0\rangle$ can be expressed as:

$$\begin{aligned}\bar{\mathcal{A}}_N &= \lim_{T\to\infty}\frac{1}{T}\int_0^T dt \sum_{ij} e^{-i(\lambda_i-\lambda_j)t} c_i^* c_j \langle\lambda_i|\mathcal{A}_N|\lambda_j\rangle \\ &= \sum_{i=0}^{|\sigma(H)|-1} |c_i|^2\langle\lambda_i|\mathcal{A}_N|\lambda_i\rangle. \end{aligned}\tag{4.40}$$

We now describe a quantum $O(\text{poly } N)$ space protocol to estimate $\bar{\mathcal{A}}_N$. An outline of our algorithm is shown below:

$$
\begin{aligned}
|0^a\rangle \left(|\psi_0\rangle|0^{3(m+1)+\lceil 2\log_2 \epsilon_2^{-1}\rceil}\rangle\right) =& |0^a\rangle \left(\sum c_i|\lambda_i\rangle|0^{3(m+1)+\lceil 2\log_2 \epsilon_2^{-1}\rceil}\rangle\right) \\
\xrightarrow{\text{Phase estimation of } \tilde{U}_{\epsilon_2}} & |0^a\rangle \left(\sum c_i|\lambda_i\rangle|\tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})}\rangle + |\zeta\rangle\right) \\
\xrightarrow{\text{Block encoding } \frac{1}{4}\sqrt{1-\mathcal{A}_N/2dN}} & |0^a\rangle \left(\sum \frac{c_i}{4}\sqrt{1-\mathcal{A}_N/2dN}|\lambda_i\rangle|\tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})}\rangle + |\zeta'\rangle\right) + |\xi\rangle \\
\xrightarrow{\text{Measure first register}} & \Pr[0^a] \approx \frac{1}{16}\sum_i |c_i|^2 \langle\lambda_i|(1-\mathcal{A}_N/2dN)|\lambda_i\rangle = \frac{1-\bar{\mathcal{A}}_N/2dN}{16}.
\end{aligned}
\tag{4.41}
$$

First, we prepare the state $|0^a\rangle \otimes \left(|\psi_0\rangle|0^{3(m+1)+\lceil 2\log_2 \epsilon_2^{-1}\rceil}\rangle\right)$. The first register has $a = \lceil N\log_2 d\rceil + 5 = O(N)$ qubits. Since $m = O(\text{poly } N)$, this state can be prepared in $O(\text{poly } N)$ space. For convenience, we use a qubit-based algorithm rather than a qudit based algorithm, where each qudit of the original state is replaced with $\lceil \log d\rceil$ qubits. We designate the first a qubits, which will act as a control, as register 1, and the remaining register as register 2.

Next, we perform phase estimation on register 2 for the operator $\tilde{U}_{\epsilon_2}$, taking it to m bits of precision and with 2-norm error in the approximation register of $\epsilon_2 2^{-(m+1)}$. The resulting state at the end of this step is:

$$
|0^a\rangle \left(\sum c_i|\lambda_i\rangle|\tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})}\rangle + |\zeta\rangle\right), \tag{4.42}
$$

where $|\tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})}\rangle$ is as above, and $|\zeta\rangle$, the error from performing the QPE on $\tilde{U}_{\epsilon_1}$ instead of e^{iH}, satisfies $\||\zeta\rangle\| \leq \epsilon_1$. We are once again able to give an efficient classical description of this phase estimation.

We now apply a block encoding of $\sqrt{1-\mathcal{A}_N/2dN}/4$ to our state. By Lemma 4.27, we

can do so to ϵ_3 error using $O(\operatorname{poly}\log N \operatorname{poly}\log \epsilon_3^{-1})$ ancillas. The resulting state is:

$$|0^a\rangle \left(\sum \frac{c_i}{4} \sqrt{1 - \mathcal{A}_N/2dN} |\lambda_i\rangle |\tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})}\rangle + |\zeta'\rangle \right) + |\xi\rangle \tag{4.43}$$

Where $(\langle 0^a| \otimes I)|\xi\rangle = 0$, and $\||\zeta'\rangle\| \leq \epsilon_1 + \epsilon_3$.

Thus, we see that if we measure the first register, we will obtain 0^a with probability

$$\begin{aligned}\Pr[0^a] = \sum_{i,j} \frac{c_i^* c_j}{16} \langle \lambda_i | \left(1 - \mathcal{A}_N/2dN\right) |\lambda_j\rangle \langle \tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})} | \tilde{\lambda}_j^{(m,\epsilon_2 2^{-(m+1)})}\rangle \\ + \sum_i \left[\frac{c_i}{4} \left(\langle \zeta' | \sqrt{1 - \mathcal{A}_N/2dN} \otimes I \right) |\lambda_i\rangle |\tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})}\rangle + \frac{c_i^*}{4} \langle \lambda_i | \langle \tilde{\lambda}_i^{(m,\epsilon_2 2^{-(m+1)})} | \left(\sqrt{\mathcal{A}_N} \otimes I |\zeta'\rangle \right) \right],\end{aligned} \tag{4.44}$$

and hence,

$$\left| \Pr[0^a] - \frac{1 - \frac{\bar{\mathcal{A}}_N}{2dN}}{16} \right| \leq \frac{\epsilon_2}{16} + \frac{1}{2}(\epsilon_1 + \epsilon_3). \tag{4.45}$$

Thus, by Hoeffding's inequality, for any $\epsilon_4 > 0$, it suffices to repeat this measurement $\lceil \epsilon_4^{-2} \rceil$ times and take the sample average to get an $(\epsilon_2/16 + (\epsilon_1 + \epsilon_3)/2 + \epsilon_4)$-estimate of $\frac{1 - \frac{\bar{\mathcal{A}}_N}{2dN}}{16}$ with probability at least $2/3$. To get overall error of ϵ in the estimate of $|\bar{\mathcal{A}}_N|$ with that probability, it suffices to take $\epsilon_2 = \epsilon/8dN$, $\epsilon_4 = \epsilon/128dN$, and $\epsilon_1, \epsilon_3 = \epsilon/64dN$. Thus, we see that to repeat this measurement, we need in total

$$O\big(\epsilon^{-2} \left(\operatorname{poly} N + \operatorname{poly}\log N\epsilon^{-1} + \operatorname{poly}\log N \operatorname{poly}\log N\epsilon^{-1} \right)\big) = O(\epsilon^{-2} \operatorname{poly}\log \epsilon^{-1} \operatorname{poly} N) \tag{4.46}$$

space. □

Finally, this algorithm allows us to decide FSRelax. Hence, we have

Lemma 4.29. FSRelax $\in$ PSPACE.

Proof. Lemma 4.28 allows us to construct a probabilistic quantum polynomial space algorithm to decide FSRelax. The algorithm in Lemma 4.28 relies on postselection. We can use Lemma 4.28 to probabilistically compute Γ such that $|\Gamma - \bar{\mathcal{A}}_N| \leq \epsilon'$ for some $\epsilon' < \epsilon$. In the event of the correct postselection outcomes occurring, since we have the promise that either $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$ or $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$, one of the two following cases holds.

1. If $|\Gamma - A^*| \leq \epsilon + \epsilon'$ then $|\bar{\mathcal{A}}_N - A^*| \leq \epsilon$.

2. Otherwise,if $|\Gamma - A^*| \geq c\epsilon - \epsilon'$, then $|\bar{\mathcal{A}}_N - A^*| \geq c\epsilon$.

Hence if our algorithm postselects correctly, FSRelax can be decided by a polynomial space quantum algorithm that computes $|\Gamma - A^*|$. On the other hand, if our algorithm does not postselect correctly, we can simply return YES or NO at random. For YES instances of FSRelax, we accept with probability strictly greater than $1/2$, while for NO instances, we accept with probability strictly less than $1/2$. Hence, FSRelax $\in$ PrQSPACE. By Watrous [Wat99], PrQSPACE $\subseteq$ PSPACE. Therefore, FSRelax $\in PSPACE$. □

Finally, our main result Theorem 4.5 follows immediately from Lemma 4.29 and Theorem 4.25.

4.5 FSTherm(MC) and FSTherm(Gibbs) are contained in PSPACE

In this section we'll show that FSTherm(MC) and FSTherm(Gibbs) are contained in PSPACE. We will first show a polynomial quantum space algorithm to compute $\mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{MC}}]$.

Along with the polynomial space algorithm to determine FSRelax, this yields a polynomial quantum space algorithm to decide FSTherm(MC) by reducing any instance of FSTherm(MC) to an instance of FSRelax. Similarly, using a polynomial space quantum algorithm to compute $\mathrm{tr}[\mathcal{A}_N \rho_{\mathsf{MC}}]$, we show that FSTherm(Gibbs) $\in$ PSPACE. Throughout this section, we restrict to systems of qubits rather than qudits, in order to make use of qubit block-encoding results. However, this containment result is general, since a system of a k-local Hamiltonian on N qudits of local dimension d can be described as a system of a $k \log \lceil d \rceil$-local Hamiltonian on $N \log \lceil d \rceil$ qubits.

First, we prove some preliminary lemmas about block-encodings, which we will use to prepare an approximation of the microcanonical state.

Proposition 4.30. *Let H be a Hermitian matrix on N qubits, and $\alpha \geq \|H\|$. Let the eigenvectors and eigenvalues of H be $\{|\lambda_i\rangle, \lambda_i\}$. For a polynomial p, let U be an $(N+m)$-qubit, $(\alpha, m, 0)$-block encoding of $p(H/\alpha)$. Let $|\psi\rangle := \frac{1}{\sqrt{2^N}} \sum_i |\lambda_i\rangle_a |\lambda_i\rangle_b$, where N qubits form the main computational register a, and N qubits form an auxiliary register b. Then, U can be used to prepare the state*

$$|\psi_{\mathrm{poly}}(p)\rangle := \frac{p(H/\alpha)_a |\psi\rangle}{\|p(H/\alpha)_a |\psi\rangle\|}. \tag{4.47}$$

such that

$$\rho_{\mathrm{poly}}(p) := \mathrm{tr}_b[|\psi_{\mathrm{poly}}(p)\rangle\langle\psi_{\mathrm{poly}}(p)|] = \sum_i \frac{(p(\lambda_i/\alpha))^2}{\sum_j (p(\lambda_i/\alpha))^2} |\lambda_i\rangle\langle\lambda_i| \tag{4.48}$$

Proof. Let c be the register of the m block encoding ancilla qubits. U acts non-trivially on registers a and c. By Definition 4.11, U prepares the state

$$U|\psi\rangle = p(H/\alpha)_a |\psi\rangle |0^m\rangle_c + |\phi^\perp\rangle, \tag{4.49}$$

where $\langle 0^m|_c|\phi^\perp\rangle = 0$, and

$$\begin{aligned} p(H/\alpha)_a|\psi\rangle &= \frac{1}{\sqrt{2^n}}\sum_i \frac{1}{\sqrt{2^N}} p(H/\alpha)_a|\lambda_i\rangle_a|\lambda_i\rangle_b \\ &= \frac{1}{\sqrt{2^N}}\sum_i p(\lambda_i/\alpha)|\lambda_i\rangle_a|\lambda_i\rangle_b. \end{aligned} \tag{4.50}$$

The state obtained after postselecting on the correct block-encoding being performed is

$$\begin{aligned} |\psi_{\mathrm{poly}}(p)\rangle &\coloneqq \frac{p(H'/\alpha)_a|\psi\rangle}{\|p(H'/\alpha)_a|\psi\rangle\|} \\ &= \frac{1}{\sqrt{\sum_j (p(\lambda_i/\alpha))^2}}\sum_i p(\lambda_i/\alpha)|\lambda_i\rangle_a|\lambda_i\rangle_b. \end{aligned} \tag{4.51}$$

It follows that

$$\rho_p \coloneqq \mathrm{tr}_b[|\psi_{\mathrm{poly}}(p)\rangle\langle\psi_{\mathrm{poly}}(p)|] = \sum_i \frac{(p(\lambda_i/\alpha))^2}{\sum_j (p(\lambda_i/\alpha))^2}|\lambda_i\rangle\langle\lambda_i|. \tag{4.52}$$

□

We now show that there exists a suitable choice of polynomial p such that ρ_p approximates the microcanonical state.

Lemma 4.31. *Let H be a Hermitian matrix on N qubits, and $\alpha \geq \|H\|$. Let $P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}$ be defined as in Lemma 4.15. Then, for any $\epsilon_p \in (0,1]$, with $\eta = \frac{\epsilon_p e^{-\pi\alpha^2/w^2}}{9}$, $P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}$ has degree $O(\frac{\alpha}{w}(\log 1/\epsilon + \pi\alpha^2/w^2))$ and*

$$\|\rho_{\mathrm{poly}}(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}) - \rho_{\mathsf{MC}}(H,0)\|_1 \leq \epsilon_p. \tag{4.53}$$

Proof. First, we note that by Lemma 4.15 there exists a polynomial approximation $p(x)$ of the Gaussian function such that

$$|P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)} - e^{-\pi\alpha^2 x^2/(2w^2)}| \leq \eta \ \ \forall\, x \in [-1,1] \tag{4.54}$$

where $P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}$ has degree $\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}) = O(\frac{\alpha}{w}\log 1/\eta)$. Note that

$$\rho_{\mathrm{poly}}(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}) = \sum_i \frac{(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}(\lambda_i/\alpha))^2}{\sum_j (P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}(\lambda_i/\alpha))^2}|\lambda_i\rangle\langle\lambda_i|, \ \ \rho_{\mathsf{MC}}(H,0) = \sum_i \frac{e^{-\pi\lambda_i^2/w^2}}{\sum_j e^{-\pi\lambda_j^2/w^2}}|\lambda_i\rangle\langle\lambda_i| \tag{4.55}$$

Let $S \coloneqq \sum_i e^{-\pi\lambda_i^2/w^2}$ and $\Delta \coloneqq \sum_i (p(\lambda_i/\alpha))^2 - e^{-\pi\lambda_i^2/w^2}$. Then,

$$\begin{aligned} |(p(\lambda_i/\alpha))^2 - e^{-\pi\lambda_i^2/w^2}| &\leq \eta|p(\lambda_i/\alpha) + e^{-\pi\lambda_i^2/(2w^2)}| \\ &\leq \eta(2+\eta) \\ \implies |\Delta| &\leq \eta(2+\eta)2^N. \end{aligned} \tag{4.56}$$

Let $\epsilon_\eta \coloneqq \eta(2+\eta)$. We thus have

$$\begin{aligned} \|\rho_{\mathrm{poly}}(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}) - \rho_{\mathsf{MC}}(H,0)\|_1 &= \|\sum_i \left(\frac{(p(\lambda_i/\alpha))^2}{S+\Delta} - \frac{e^{-\pi\lambda_i^2/w^2}}{S}\right)|\lambda_i\rangle\langle\lambda_i|\|_1 \\ &= \sum_i \left|\frac{S(p(\lambda_i/\alpha))^2}{S(S+\Delta)} - \frac{(S+\Delta)e^{-\pi\lambda_i^2/w^2}}{S(S+\Delta)}\right| \\ &\leq \sum_i \left(\frac{\epsilon_\eta}{|S+\Delta|} + \left|\frac{\Delta e^{-\pi\lambda_i^2/w^2}}{S(S+\Delta)}\right|\right) \\ &\leq \frac{\epsilon_\eta 2^N + |\Delta|}{|S+\Delta|} \\ &\leq \frac{2\epsilon_\eta 2^N}{|S+\Delta|}. \end{aligned} \tag{4.57}$$

Since $\lambda_i/\alpha \in [-1,1]$, $S \geq e^{-\pi\alpha^2/w^2}2^N$. Therefore,

$$\|\rho_{\text{poly}}(P^{(\eta)}_{\text{Gauss}(w/\alpha)}) - \rho_{\mathsf{MC}}(H,0)\|_1 \leq \frac{2\epsilon_\eta}{e^{-\pi\alpha^2/w^2} - \epsilon_\eta}, \tag{4.58}$$

Selecting $\eta \leq \frac{\epsilon_p e^{-\pi\alpha^2/w^2}}{9} < 1$, we have

$$\begin{aligned} \epsilon_\eta &\leq \frac{\epsilon e^{-\pi\alpha^2/w^2}}{3} \\ \implies \|\rho_{\text{poly}}(P^{(\eta)}_{\text{Gauss}(w/\alpha)}) - \rho_{\mathsf{MC}}(H,0)\|_1 &\leq \epsilon_p. \end{aligned} \tag{4.59}$$

The degree of p is $O(\frac{\alpha}{w}\log 1/\eta) = O(\frac{\alpha}{w}(\log 1/\epsilon + \pi\alpha^2/w^2))$, which completes the proof. □

Lemma 4.32. *Let H be a k-local Hamiltonian on N qubits, $\alpha \leq \|H\|$, and let the eigenvectors and eigenvalues of H be $\{|\lambda_i\rangle, \lambda_i\}$. For any $\epsilon_p \in (0,1]$, let $P^{(\eta)}_{\text{Gauss}(w/\alpha)}$ be chosen as in Lemma 4.31. Let U be an $(N+m)$-qubit, (α, m, ϵ')-block encoding of $P^{(\eta)}_{\text{Gauss}(w/\alpha)}(H/\alpha)$. Then, a single application of U (postselecting on the correct block-encoding outcome) can be used to prepare a state ρ' such that*

$$\|\rho' - \rho_{\mathsf{MC}}(H,0)\|_1 \leq \sqrt{8\epsilon'} + \epsilon_p. \tag{4.60}$$

Proof. As in Lemma 4.31, let $|\psi\rangle := \sum_i \frac{1}{\sqrt{2^N}}|\lambda_i\rangle_a|\lambda_i\rangle_b$, where N qudits form the main computational register a, N qudits form an auxiliary register b, and $|\lambda_i\rangle$ are the eigenstates of H/α. $|\psi\rangle$ is the maximally entangled state represented in the basis of eigenstates of H/α. We also define a register c for the n_c block encoding ancilla qudits. U acts on all three of these registers and by

definition of the block-encoding,

$$\|\langle 0^{n_c}|_c U|0^{n_c}\rangle_c - p(H/\alpha)\| \leq \epsilon'. \tag{4.61}$$

Let

$$|\psi_p\rangle \coloneqq \frac{p(H/\alpha)_a|\psi\rangle}{\|p(H/\alpha)_a|\psi\rangle\|}. \tag{4.62}$$

Then, as in Proposition 4.30

$$\rho_{\text{poly}} = \text{tr}_b[|\psi_p\rangle\langle\psi_p|]. \tag{4.63}$$

Let $M \coloneqq \langle 0^{n_c}|_c U|0^{n_c}\rangle_c$. Then, let the state obtained after postselecting on the correct block-encoding being performed be

$$|\phi\rangle = \frac{M|\psi\rangle}{\|M|\psi\rangle\|}. \tag{4.64}$$

Further, let

$$\rho' \coloneqq \text{tr}_b[|\phi\rangle\langle\phi|]. \tag{4.65}$$

Then, by Proposition 9 of Childs et al. [CKS17],

$$\||\psi_p\rangle - |\phi\rangle\| \leq \epsilon', \tag{4.66}$$

By monotonicity of the trace distance under the partial trace,

$$\|\text{tr}_b[|\phi\rangle\langle\phi|] - \text{tr}_b[|\psi_p\rangle\langle\psi_p|]\|_1 \leq \||\phi\rangle\langle\phi| - |\psi_p\rangle\langle\psi_p|\|_1. \tag{4.67}$$

Using this, and the relation between the fidelity and the trace distance, we obtain

$$\begin{aligned} \|\rho' - \rho_{\text{poly}}\|_1 &\leq 2\sqrt{2\epsilon' - \epsilon'^2} \\ &\leq \sqrt{8\epsilon'}. \end{aligned} \tag{4.68}$$

Combining this with Lemma 4.31 using the triangle inequality, we obtain

$$\|\rho' - \rho_{\mathsf{MC}}(H, 0)\|_1 \leq \sqrt{8\epsilon'} + \epsilon_p. \tag{4.69}$$

□

Finally, we have the required ingredients to prepare the microcanonical state.

Lemma 4.33. *For any k-local Hamiltonian H on N qubits, and energy E, there is a quantum algorithm using $O(\text{poly}\, N, \log 1/\epsilon)$ space and $O(\text{poly}\, N, \log 1/\epsilon \exp(((\|H\| + |E|)/w)^2))$ expected time that prepares a state ρ such that*

$$\|\rho - \rho_{\mathsf{MC}}(H, E)\|_1 \leq \epsilon. \tag{4.70}$$

Proof. Our algorithm is simple. Starting in the initial state $|\psi\rangle := \frac{1}{\sqrt{2^s}} \sum_i |\lambda_i\rangle_a |\lambda_i\rangle_b$, we apply a block encoding such that the resultant state, after postselecting on the correct block encoding being performed and tracing out the subsystem b, is a good approximation of the microcanonical state. It remains to be shown that such a block encoding can be constructed, and to analyze the approximation error as well as the time and space requirements.

First, we show how the block encoding is constructed. Since each term of H is k-local

and there are $O(\operatorname{poly} n)$ terms in H, $H' := H - EI$ is $O(2^k \operatorname{poly} n)$-sparse. We can select $\alpha \geq \|H\| + |E|$ so that the eigenvalues of H'/α lie in $[-1, 1]$. $\alpha = \operatorname{poly} n$, so by Lemma 4.12, H' has an $(\operatorname{poly} n, \operatorname{poly} n, \epsilon_H)$-block-encoding that can be implemented by a circuit with $O(\operatorname{poly} n, \operatorname{poly} \log \frac{1}{\epsilon_H})$ gates and $O(\operatorname{poly} n, \operatorname{poly} \log \frac{1}{\epsilon_H})$ space.

We make use of the polynomial approximation for a Gaussian function $P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}$ (Lemma 4.15) with degree $\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)})$ which we will determine later. Along with Lemma 4.13, this implies there is a circuit $\tilde{U}_\delta$ with $O(\operatorname{poly} n, \operatorname{poly} \log \frac{1}{\epsilon_H})$ gates that implements a $(1, \operatorname{poly} n, 4 \deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}) \sqrt{\epsilon_H/\alpha} + \delta)$ block-encoding of $P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}(H'/\alpha)$, and has a classical description computable in time $O(\operatorname{poly} n, \deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}), \log 1/\delta)$.

The above construction makes use of parameters η, ϵ_H, δ. We would like to choose these parameters such that we approximate the target state well, and the postselection probability is not too small. For the first condition, we apply Lemma 4.32 with $\epsilon' = 4 \deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}) \sqrt{\epsilon_H/\alpha} + \delta$, we have

$$\|\rho' - \rho_{\mathsf{MC}}(H, 0)\|_1 \leq \left(\sqrt{32 \deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}) \sqrt{\epsilon_H/\alpha} + 8\delta} \right) + \epsilon_p. \tag{4.71}$$

We will solve for parameters such that the following condition is fulfilled:

$$\|\rho' - \rho_{\mathsf{MC}}\|_1 \leq \epsilon. \tag{4.72}$$

For our second condition, we require that our algorithm doesn't take too long. The time taken is determined by the probability of correctly postselecting, which we now estimate. Let $M := \langle 0^{n_c}|_c \tilde{U}_\delta |0^{n_c}\rangle_c$. The probability of postselecting to the correct state is $\|M|\psi\rangle\|^2$. Observe

that

$$\begin{aligned}\|M|\psi\rangle\| &\geq \|P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}(H'/\alpha)|\psi\rangle\| - \|(M - P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}(H'/\alpha))|\psi\rangle\| \\ &\geq \|P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}(H'/\alpha)|\psi\rangle\| - 4\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)})\sqrt{\epsilon_H/\alpha} - \delta \\ &\geq e^{-\pi\alpha^2/(2w^2)} - \eta - 4\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)})\sqrt{\epsilon_H/\alpha} - \delta. \end{aligned} \tag{4.73}$$

In order for the probability of correct postselection to be lower bounded, require that the probability of correct postselection remains of the order of its leading term, i.e.,

$$\|M|\psi\rangle\| = \Omega(\exp(-\pi\alpha^2/(2w^2))). \tag{4.74}$$

We now solve for η, ϵ_H, δ such that Eq. (4.72) and Eq. (4.74) are fulfilled.

First, to partially satisfy Eq. (4.72), we select $\epsilon_p = \epsilon/2$. This implies $\eta \leq \frac{\epsilon e^{-\pi\alpha^2/w^2}}{18}$. Eq. (4.72) is then satisfied if $\sqrt{8\delta} < \epsilon/4$ and $\sqrt{32\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)})\sqrt{\epsilon_H/\alpha}} < \epsilon/4$.

Now, in order to satisfy Eq. (4.74), we impose the constraint that $\eta, 4\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)})\sqrt{\epsilon_H/\alpha}, \delta$ are each less than $\frac{1}{4}e^{-\pi\alpha^2/(2w^2)}$. By our choice of ϵ_p, we have already obtained a sufficient condition for η.

To simultaneously fulfill $\sqrt{8\delta} < \epsilon/4$ and $\delta < \frac{1}{4}e^{-\pi\alpha^2/(2w^2)}$, we select $\delta < \epsilon^2 e^{-\pi\alpha^2/(2w^2)}/32$.

We make use of the fact that $\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)}) = \gamma\left(\frac{\alpha}{w}(\log 2/\epsilon + \pi\alpha^2/w^2)\right)$ for some constant γ. To simultaneously fulfill $\sqrt{32\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)})\sqrt{\epsilon_H/\alpha}} < \epsilon/4$ and $4\deg(P^{(\eta)}_{\mathrm{Gauss}(w/\alpha)})\sqrt{\epsilon_H/\alpha} < \frac{1}{4}e^{-\pi\alpha^2/(2w^2)}$, we select

$$\epsilon_H \leq \alpha\left(\frac{\epsilon e^{-\pi\alpha^2/(2w^2)}}{128\gamma\left((\alpha/w)(\log 2/\epsilon + \pi\alpha^2/w^2)\right)}\right)^4. \tag{4.75}$$

With the chosen values of η, ϵ_H, δ, Eq. (4.72) and Eq. (4.74) are satisfied. We now analyze the time and space requirements of the algorithm. The probability of correct postselection is $\|M|\psi\rangle\|^2$. Hence, in order to postselect onto the target state ρ', we require an expected $O(e^{\pi\alpha^2/w^2})$ repetitions of the block encoding circuit. We do not need any additional space despite an expected total exponential depth since we are repeating the same circuit until correct postselection. Further, the depth of the block-encoding circuit as well as the classical computation time to obtain its description remain polynomial since $\log \frac{1}{\epsilon_H}, \log \frac{1}{\delta} = O(\log 1/\epsilon(\alpha/w)^2)$.

The depth of the block-encoding circuit scales as

$$O(\left(\frac{\alpha}{w}\right) \log 1/\eta \times \operatorname{poly} \log 1/\epsilon_H) = O(\operatorname{poly} N, \log 1/\epsilon, 1/w). \tag{4.76}$$

Therefore, to prepare ρ we require $O(e^{\pi\alpha^2/w^2})$ expected executions of a $O(\operatorname{poly} N, \log 1/\epsilon, 1/w)$ depth circuit, which can be performed in $O(\operatorname{poly} N, \log 1/\epsilon, 1/w)$ space. □

This algorithm allows us to estimate observables of the microcanonical state.

Corollary 4.34. *The algorithm in Lemma 4.33 can be used to estimate* $\widehat{\mathcal{A}_N}$ *such that* $|\widehat{\mathcal{A}_N} - \operatorname{tr}[\mathcal{A}_N \rho_{MC}]| \leq \epsilon$ *using* $O(\operatorname{poly} N, \log 1/\epsilon)$ *space and* $O(\operatorname{poly} N, \log 1/\epsilon \exp(((\|H\| + |E|)/w)^2))$ *expected time.*

Proof. Note that by the triangle inequality, $\|A_N\| \leq \max_i \|A_i\| = c$ for some constant c. Let ρ

be the state as prepared in Lemma 4.33, to precision ϵ/c. Applying Holder's inequality,

$$\begin{aligned} |\operatorname{tr}[\mathcal{A}_N\rho - \mathcal{A}_N\rho_{\mathsf{MC}}(H,E)]| &\leq \|\mathcal{A}_N(\rho - \rho_{\mathsf{MC}}(H,E))\|_1 \\ &\leq \|\rho - \rho_{\mathsf{MC}}(H,E)\|_1\|\mathcal{A}_N\| \\ &\leq \epsilon. \end{aligned} \tag{4.77}$$

Hence by preparing ρ and measuring $\mathcal{A}_N$, we can estimate $\operatorname{tr}[\mathcal{A}_N\rho_{\mathsf{MC}}]$ to arbitrary constant precision. $\mathcal{A}_N$ is a sum of N single site (and hence commuting) observables. A single measurement of $\mathcal{A}_N$ can be performed by preparing a sample of ρ and measuring each of the A_i observables sequentially, adding them, and dividing by N. By Hoeffding's inequality, to estimate $\operatorname{tr}[\mathcal{A}_N\rho_{\mathsf{MC}}]$ to precision ϵ with probability at least $2/3$ requires $O(1/\epsilon^2)$ samples. Hence there is a a probabilistic polynomial space quantum algorithm to estimate $\operatorname{tr}[\mathcal{A}_N\rho_{\mathsf{MC}}]$ to constant precision. We note however, that this algorithm requires postselection, which may only succeed with inverse-exponentially small probability when $w = O(\operatorname{poly} N)$. □

Finally, we can use this algorithm to prove one of our main results.

Theorem 4.35. $\mathsf{FSTherm}(\mathsf{MC}) \in \mathsf{PSPACE}$.

Proof. Using the algorithm from Corollary 4.34 along with the probabilistic polynomial space quantum algorithm to estimate $\bar{\mathcal{A}}_N$ to constant precision Lemma 4.28, when $w = \Omega(\frac{1}{\operatorname{poly} N})$, $\mathsf{FSTherm}(\mathsf{MC})$ can be decided by a polynomial-space quantum algorithm that simply compares the two values, if our algorithm postselects correctly, which occurs with non-zero probability. On the other hand, in the event our algorithm does not postselect correctly, we can Accept or Reject with uniform probability. Hence, there is a quantum algorithm that such that for YES instances of

$\mathsf{FSTherm}(\mathsf{MC})$, we accept with probability strictly greater than $1/2$, while for NO instances, we accept with probability strictly less than $1/2$. Hence, $\mathsf{FSTherm}(\mathsf{MC}) \in \mathsf{PrQSPACE}$. By Watrous [Wat99], $\mathsf{PrQSPACE} \subseteq \mathsf{PSPACE}$. Therefore, $\mathsf{FSTherm}(\mathsf{MC}) \in \mathsf{PSPACE}$. □

Another statistical ensemble of interest is the Gibbs ensemble $\rho_{\mathsf{G}} = e^{-\beta H}/\operatorname{tr}[e^{-\beta H}]$, representing the state of a system in contact with a bath of temperature $1/\beta$. An alternative to $\mathsf{FSTherm}(\mathsf{MC})$ is the problem of deciding whether a system thermalizes to the *Gibbs* ensemble, i.e., whether the long-time average of an observable converges to the value corresponding to the Gibbs ensemble. Here, we note that the Gibbs state can be prepared probabilistically in polynomial quantum space, and hence this decision problem is also contained in PSPACE.

Lemma 4.36. *There is a quantum algorithm using $O(\operatorname{poly} N, \log 1/\epsilon)$ space and $O(\exp(N))$ time that computes $\widetilde{\mathcal{A}_N}$ such that*

$$|\operatorname{tr}[\mathcal{A}_N \rho_G] - \widetilde{\mathcal{A}_N}| \leq \epsilon. \tag{4.78}$$

Proof. The Gibbs state $\rho_{\mathsf{G}} = \frac{e^{-\beta H}}{\operatorname{tr}[e^{-\beta H}]}$ can be ϵ-approximated with a block-encoding of circuit depth $O(\sqrt{\beta} \log 1/\epsilon)$ and $O(\sqrt{d^N/Z})$ amplitude amplification steps [Gil+19; Ape+20; CS16]. Note that $Z = \operatorname{tr}[e^{-\beta H}] \geq d^N e^{-\beta \|H\|} = \Omega(1/e^{\operatorname{poly} N})$, so estimating $\operatorname{tr}[\rho_{\mathsf{G}} \mathcal{A}_N]$ to constant precision only requires $O(\exp(N))$ repetitions of a $O(\operatorname{poly} N)$-depth circuit, which can be done in polynomial quantum space. □

Since $\mathsf{PrQSPACE} \subseteq \mathsf{PSPACE}$, Theorem 4.7 follows immediately from Lemma 4.36.

4.6 Hardness of FSTherm(MC) and FSTherm(Gibbs)

In this subsection we will show that FSTherm(MC) is as hard as FSHalt under *quantum* polynomial time reductions. We prove this by showing that given an instance of FSHalt, we may use a poynomial time quantum algorithm to construct an instance of FSTherm(MC). We note that since we make use of a quantum polynomial time reduction rather than a classical one, our reduction differs from the standard formal complexity theoretic reduction. We modify the FSRelax hardness construction such that we have $A^* = \mathrm{tr}[\rho_{\mathsf{MC}}\mathcal{A}_N]$, thus showing that there exist observables such that determining whether the system thermalizes is as hard as a PSPACE-complete problem. We describe a high-level overview of our approach:

1. We pad the end of the Turing machine even more in the case of halting, so that when it halts, $\bar{\mathcal{A}}_N = (1 + O(1/N))(1 - \alpha)\langle a_2|A|a_2\rangle$. This is covered in Section 4.6.1.

2. We double the size of the local Hilbert space, and modify A_i by introducing parameters such that the two halves of the Hilbert space have different, tunable values of $\langle A_i\rangle$. We then modify the Hamiltonian so that, in the microcanonical state, both halves of the Hilbert space are represented equally. This construction allows us to tune the microcanonical expectation of $\mathcal{A}_N$. This is covered in Section 4.6.2.

3. Finally, we show that a polynomial time quantum algorithm can be used to tune these parameters such that in the case of halting, $\bar{\mathcal{A}}_N$ is equal to the microcanonical average of $\mathcal{A}_N$, and both are separated from 0, the non-halting value of $\mathcal{A}_N$. This is covered in Section 4.6.3.

Finally, in Section 4.6.4, we discuss some evidence and suggest a conjecture under which

FSTherm(Gibbs) is PSPACE-hard (and hence PSPACE-complete due to Lemma 4.36).

4.6.1 Modifying the TM

We modify the Turing Machine construction such that the values of $\bar{\mathcal{A}}_N$ on halting inputs and on non-halting inputs are sufficiently separated, so that we may perform our tuning procedure in the following sections to reduce FSRelax to FSTherm(MC). We first show how to asymptotically increase $\bar{\mathcal{A}}_N$ in the case of halting to $(1-\alpha)\langle a_2|A|a_2\rangle$, which we require for technical reasons in our tuning construction in Section 4.6.2. As discussed in Section 4.3.1, TM1 performs the actual computation of a URTM, TM2 flips A cells if TM1 halts, and TM3 runs TM1 in reverse. Since in the case of halting, the A cells are flipped by TM2 and remain flipped for the duration of TM3, the long-time average of $\mathcal{A}_N$ is determined by the duration of TM2 and TM3. We can thus modify this average by changing the time spent in TM3. We achieve this by modifying TM3 to first run a buffer which performs a full loop around the tape for each element of the tape (i.e., taking $N(N+1)$ steps). Then, we continue as before with TM3 running TM1 in reverse, but with an additional buffer that iterates the finite control around the entire loop until it returns to its initial location in between each (computational) step of TM3. Since a full loop around the tape takes time N, we thus increase the time spent in the computational part of TM3 by a factor of $N+1$.

To ensure the correct TM action, we add to the tape alphabet of TM3 such that every symbol in the tape alphabet has a "place-marker" copy which will be what the TM initially writes to the tape. This copy lets the finite-control know when it has returned to its most recent write cell. To facilitate the traversal, we will also add a copy of every finite-control state, which the update rules

will just shift through the ring until it hits the "place-marker" tape-symbols. For the first buffer, we need to add another state of the head corresponding to the specific buffer, and two symbols to the tape alphabet to help keep track of the progress of the N loops around the tape. Specifically we will initialize the two symbols, pushing one of the symbols in one direction after every loop, and only finishing when the symbols come in contact again. Thus, in the case of halting, this additional looping increases the total run-time to $(N+2)T_h + (N+1)N$. The first buffer ensures that this run-time is large even when $T_h < N$.

Lemma 4.37. *With the above modified Turing Machine construction, If y encodes a Turing Machine that either halts in time T_h or never halts, then there exists $\delta \in (0, N^{-1}]$ such that*

$$\bar{\mathcal{A}}_N = \begin{cases} (1-\alpha)\left[1-\delta\right]\langle a_2|A|a_2\rangle & \textit{halting} \\ \\ 0 & \textit{non-halting}. \end{cases} \tag{4.79}$$

Proof. The proof proceeds along exactly the same lines as the proof of Lemma 4.24. Revisiting Eq. (4.29), with the modification that $T = (N+2)T_h + (N+1)N$, we have that

$$\begin{aligned} \bar{\mathcal{A}}_N &= (1-\alpha)\frac{T - (T_h + N/2 + 1/2)}{T+1} \\ &= (1-\alpha)\frac{(N+2)T_h + (N+1)N - (T_h + N/2 + 1/2)}{(N+2)T_h + (N+1)N + 1}\langle a_2|A|a_2\rangle \\ &= (1-\alpha)\left[1 - \frac{T_h + N/2 + 3/2}{(N+2)T_h + (N+1)N + 1}\right]\langle a_2|A|a_2\rangle \\ &= (1-\alpha)\left[1-\delta\right]\langle a_2|A|a_2\rangle, \end{aligned} \tag{4.80}$$

where $\delta := \frac{T_h + N/2 + 3/2}{(N+2)T_h + (N+1)N + 1}$, and $0 < \delta \leq 2N^{-1}$. The non-halting case will still have $\bar{\mathcal{A}}_N = 0$. □

4.6.2 Modifying H, $\mathcal{H}$, and A

In addition to this modification to the Turing Machine, we will also modify the Hilbert space and Hamiltonian. The goal of this modification is to:

- Push the microcanonical expectation value of $\mathcal{A}_N$ to the middle of A's spectrum.
- Give us control of $\bar{\mathcal{A}}_N$ in the halting case in such a way that we are guaranteed that it crosses the microcanonical expectation away from the non-halting case's value (zero).

To do these, we will replace our local Hilbert space $\mathcal{H}$ with a new Hilbert space $\mathcal{H}_{tune} := \mathcal{H} \oplus \mathcal{H}$ of doubled dimension. We will denote the projection onto $\mathcal{H} \oplus \mathbf{0}$ by Π and the projection onto $\mathbf{0} \oplus \mathcal{H}$ by Π'. Denote the basis of $\mathcal{H} \oplus \mathbf{0}$ by $\{|a\rangle \ : \ a \in Q \cup \Gamma\}$ and that of $\mathbf{0} \oplus \mathcal{H}$ by $\{|a'\rangle \ : \ a \in Q \cup \Gamma\}$. Let

$$H_{TM} \coloneqq \sum_{abcd \in \mathcal{T}} |cd\rangle\langle ab| + h.c, \tag{4.81}$$

represent Turing-machine evolution on $\mathcal{H} \oplus \mathbf{0}$, and

$$H'_{TM} \coloneqq \sum_{abcd \in \mathcal{T}} |c'd'\rangle\langle a'b'| + h.c, \tag{4.82}$$

represent Turing machine evolution on $\mathbf{0} \oplus \mathcal{H}$. We then let $H \coloneqq H_{TM} + H'_{TM}$. Note that $[H_{TM}, H'_{TM}] = 0$. We also define

$$F \coloneqq \sum_{a \in Q \cup \Gamma} |a'\rangle\langle a| + h.c. \tag{4.83}$$

Next, we slightly modify A_i. Let $A_i \coloneqq p|a_2\rangle\langle a_2|_i + q\Pi'_i$, where p, q are parameters we will

choose. With these modifications, we compute the expectation of $\mathcal{A}_N$ for the microcanonical ensemble and Gibbs ensembles as special cases of any ensemble which is analytic in H.

Lemma 4.38. *Consider any $\rho \propto f(H)$ where f is analytic and ρ is unit trace. Then,*

$$\operatorname{tr}[\mathcal{A}_N \rho] = \frac{q}{2} + p \operatorname{tr}[|a_2\rangle\langle a_2|_1 \rho]. \tag{4.84}$$

Proof. First, note that $[F, H] = 0$. As $\rho \propto f(H)$ for some analytic function $f : \mathbb{R} \to \mathbb{R}$, it follows that $[\rho_{\mathsf{MC}}, F] = 0$ by the existence of a power series expansion of f; we note that this holds for arbitrary functions f whose domain contains the spectrum of H, but this is a bit more involved to prove. Therefore, as $F^2 = I$,

$$\begin{aligned} \operatorname{tr}[\Pi_i' \rho] &= \operatorname{tr}[\Pi_i' F \rho F] \\ &= \operatorname{tr}[F \Pi_i' F \rho] \\ &= \operatorname{tr}[\Pi_i \rho] \\ &= \operatorname{tr}[(I - \Pi_i')\rho]. \end{aligned} \tag{4.85}$$

Thus, as ρ is unit-trace, $\operatorname{tr}[\Pi_i' \rho] = 1/2$. Hence, we have

$$\begin{aligned} \operatorname{tr}[\mathcal{A}_N \rho] &= \frac{p}{N} \sum_i \left(\operatorname{tr}[|a_2\rangle\langle a_2|_i \rho] + q \operatorname{tr}[\Pi_i' \rho] \right) \\ &= \frac{q}{2} + p \operatorname{tr}[|a_2\rangle\langle a_2|_1 \rho]. \end{aligned} \tag{4.86}$$

where we have used the fact that ρ is translationally invariant since H is translationally invariant, which follows from the same logic as showing that $[\rho, F] = 0$ above. □

Applying Lemma 4.38 to the microcanonical state we have

Corollary 4.39.

$$\mathrm{tr}[\mathcal{A}_N \rho_{MC}] = \frac{q}{2} + p\,\mathrm{tr}[|a_2\rangle\langle a_2|_1 \rho_{MC}]. \tag{4.87}$$

Likewise, applying Lemma 4.38 to the Gibbs state we have

Corollary 4.40.

$$\mathrm{tr}[\mathcal{A}_N \rho_{G}] = \frac{q}{2} + p\,\mathrm{tr}[|a_2\rangle\langle a_2|_1 \rho_{G}]. \tag{4.88}$$

4.6.3 A quantum polynomial time reduction from FSHalt to FSTherm(MC)

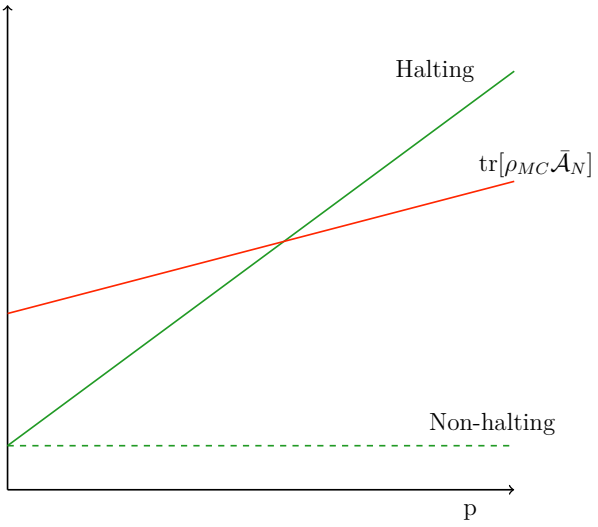


Figure 4.2: The solid green and dashed-green lines represent the time-averaged expectation values in the halting and non-halting cases respectively, as p is tuned. The red line represents the expectation values with respect to the microcanonical ensemble as p is tuned.

We now show how p can be chosen so that the long time average of $\mathcal{A}_N$, i.e., $\bar{\mathcal{A}}_N$, in the halting case and the expectation of $\mathcal{A}_N$ for the microcanonical ensemble both agree. We briefly comment here on some constraints of this construction. Our reduction relies on a quantum algorithm to prepare the microcanonical state, but the algorithm presented in Section 4.5 in general

requires $O(e^{\pi\alpha^2/w^2})$ expected executions of a $O(\text{poly } N, \log 1/\epsilon, 1/w)$ depth circuit, where α upper bounds $\|H\| + |E|$. We therefore show hardness for a restricted class of microcanonical states with width $w = \Omega(\alpha/\sqrt{\log N})$. Such a window width is still vanishingly small with respect to the range of the Hamiltonian spectrum, but allows for efficient preparation of the microcanonical state.

Lemma 4.41. *Let H be a k-local Hamiltonian with cN terms, each with norm bounded by 1. For energy E, let $\rho_{MC}(H, E)$ be the microcanonical state with window width $w = \Omega(cN/\log N)$. Then, for any ϵ there exists N_0 such that for all $N \geq N_0$ there is a quantum algorithm that takes time $O(\text{poly } N, \log \frac{1}{\epsilon})$ to compute p such that with the above Turing machine construction and observable $\mathcal{A}_N$,*

$$|\text{tr}[\rho_{MC}\mathcal{A}_N] - \bar{\mathcal{A}}_N| \leq \epsilon \tag{4.89}$$

Proof. In the above sections using Lemma 4.37 and Corollary 4.39, we showed how an instance of FSHalt can be used to construct an instance of FSRelax with parameters p, q such that:

1. In the case of non-halting, $\bar{\mathcal{A}}_N = 0$
2. In the case of halting, $\bar{\mathcal{A}}_N = (1-\alpha)\,[1-\delta]\,p.$, with $\delta < 1/N$.
3. $\text{tr}[\rho_{\text{MC}}\mathcal{A}_N] = \frac{q}{2} + \frac{p\,\text{tr}[\rho_{\text{MC}}|a_2\rangle\langle a_2|]}{2}$

In the case of halting, to tune p such that $\bar{\mathcal{A}}_N \approx \text{tr}[\rho_{\text{MC}}\mathcal{A}_N]$, we require that q is chosen such that $\text{tr}[\rho_{\text{MC}}\mathcal{A}_N] > \bar{\mathcal{A}}_N$ at $p = 0$, and $\text{tr}[\rho_{\text{MC}}\mathcal{A}_N] < \bar{\mathcal{A}}_N$ at $p = 1$. Note that as $\Pi - |a_2\rangle\langle a_2| \succeq 0$, it follows that $0 \leq \text{tr}[|a_2\rangle\langle a_2|_i \rho_{\text{MC}}] \leq \text{tr}[\Pi_i \rho_{\text{MC}}] \leq 1/2$. Using the fact that $\text{tr}[\rho_{\text{MC}}|a_2\rangle\langle a_2|_1] \leq 1/2$,

the above constraint on q is reduced to

$$\frac{q}{2} \leq (1-\alpha)[1-\delta] - \frac{1}{4}. \tag{4.90}$$

when $q = \frac{1}{2}$, for any N there exists a sufficiently small constant α such that this constraint is satisfied. Therefore, in the rest of this section we assume $q = \frac{1}{2}$. Using the polynomial time quantum algorithm described in Lemma 4.33, for constant w, we can estimate $\widetilde{\mathcal{A}_N}$ such that $|\mathrm{tr}[\rho_{\mathsf{MC}}|a_2\rangle\langle a_2|_1] - \widetilde{\mathcal{A}_N}| \leq \epsilon/2$ for constant error $\epsilon/2$. To ensure that in the case of halting, $\bar{\mathcal{A}}_N \approx \mathrm{tr}[\rho_{\mathsf{MC}}\mathcal{A}_N]$, we would like to select an appropriate p, corresponding the point of intersection of the 2 values in Fig. 4.2. With $p = \frac{1}{4(1-\alpha)-2\widetilde{\mathcal{A}_N}}$, we obtain

$$|\mathrm{tr}[\rho_{\mathsf{MC}}\mathcal{A}_N] - \bar{\mathcal{A}}_N| \leq \left|\frac{(1-\alpha)\delta}{4(1-\alpha)-2\widetilde{\mathcal{A}_N}}\right| + \frac{\epsilon}{4} \tag{4.91}$$

Since $\delta \leq N^{-1}$, there exists N_0 such that for all $N \geq N_0$, $|\mathrm{tr}[\rho_{\mathsf{MC}}\mathcal{A}_N] - \bar{\mathcal{A}}_N| \leq \epsilon$. □

We note that as we have doubled our Hilbert space dimension twice from the relaxation construction, $d = 204.$ in the above construction. Combining Theorem 4.25 and Lemma 4.41, we obtain the following theorem.

Theorem 4.42. *For any instance of* FSHalt*, there exists a Hamiltonian H, input state $|\psi_0\rangle$, and sum of local observables $\mathcal{A}_N$ for which the long time average of $\mathcal{A}_N$ equals $\mathrm{tr}[\mathcal{A}_N\rho_{\mathsf{MC}}]$ in the case of halting and equals 0 in the case of non-halting. Further, $H, |\psi_0\rangle, \mathcal{A}_N$ can be constructed by a polynomial time quantum algorithm from instances of* FSHalt*. Hence,* FSTherm(MC) *is* PSPACE*-hard under a quantum polynomial time reduction.*

Our main result Theorem 4.6 follows from Theorem 4.42 and Theorem 4.35.

4.6.4 A conjectured classical polynomial time reduction from FSHalt to FSTherm(Gibbs)

We will show evidence that FSHalt reduces to FSTherm(Gibbs) using similar techniques to Section 4.6.3. We present the outline of a classical polynomial time reduction, and state a conjecture which, if true, would imply that FSTherm(Gibbs) is PSPACE-hard.

The only notable difference from Section 4.6.2 is that we must compute appropriate β such that $\mathrm{tr}[H\rho_{\mathsf{G}}] = E \coloneqq \langle\psi_0|H|\psi_0\rangle$ efficiently classically, and then compute $\mathrm{tr}[|a_2\rangle\langle a_2|\rho_{\mathsf{G}}]$ efficiently classically in order to correctly tune the value of p (using Corollary 4.40). We can compute both of these using the result [AC21] that there is a classical polynomial time algorithm to estimate local observables of Gibbs states with finite β in 1D. However, in order to do this, we must know in advance limits on the value of $|\beta|$. In order to make use of these algorithms we require that β is constant, which we conjecture below.

In order to estimate β, we can make use of the fact that $E(\beta) = \mathrm{tr}[H\rho_{\mathsf{G}}]$ is monotonic in β. If β is guaranteed to be $O(1)$, we can use the algorithm from [AC21] to estimate $\mathrm{tr}[H\rho_{\mathsf{G}}]$ and iteratively improve our estimate of β to error ϵ in $\log 1/\epsilon$ steps classically using a binary search algorithm.

Finally, if $\beta = O(1)$, we may efficiently compute the expectation value of $\mathrm{tr}[|a_2\rangle\langle a_2|\rho_{\mathsf{G}}]$, once again using the result of Alhambra and Cirac [AC21].

How do we guarantee that $\beta = O(1)$? Observe that as β approaches $-\infty$ or ∞, ρ_{G} approaches the highest excited state or the ground state respectively. Hence, we expect that there is a range of expected energies for which β is constant, which are found away from the edges of the spectrum of H. If E is extensively gapped from the edges of the spectrum, in the thermodynamic

limit (i.e., large N limit) the energy density must be a constant distance away from the edges of the spectrum as well. This would imply that the corresponding β must be finite in the thermodynamic limit (as it is only infinite at the edges of the spectrum), and hence in the finite-sized limit we expect it to be constant. We thus conjecture that:

Conjecture 4.43. *For any 1D nearest neighbor Hamiltonian H, and product state $|\psi\rangle$, let $E = \langle\psi|H|\psi\rangle$. Let the maximum and minimum eigenvalues of H be E_{max}, E_{min} respectively. If there exists a constant ϵ such that $E_{min} + \epsilon N \leq E \leq E_{max} - \epsilon N$, and if β is such that $\mathrm{tr}[\rho_{\mathsf{G}} H] = E$, then $\beta = O(1)$*

This conjecture implies that for a Hamiltonian H used in our hardness construction, if the initial state is guaranteed to be in the middle of the spectrum, the corresponding Gibbs state has $\beta = O(1)$. By simple modifications of H (increasing the local Hilbert space dimension and applying local energy penalties), it is possible to ensure that the initial product state encoding hard instances of FSHalt has expected energy extensively separated from the edges of the spectrum. Hence, if this conjecture is true, FSHalt reduces to FSTherm(Gibbs), and FSTherm(Gibbs) is PSPACE-complete. In future work, we hope to verify this conjecture.

4.7 BQP-Hardness of FTFSRelax

4.7.1 Modifying the TM-to-Hamiltonian Construction

We return to FTFSRelax Definition 4.22 showing that it is BQP-hard following a very similar proof to the PSPACE-hardness proof of FSRelax. However, instead of encoding the problem FSHalt, we instead choose to encode a BQP-complete computation. The only major change this

requires is that TM1 is no longer a classical Turing Machine, but is replaced with a quantum Turing Machine (QTM). A QTM functions similarly to a classical TM, but where the TM head and tape configuration can be in a superposition of states [BV97]. At each time-step of the QTM evolution, the tape configuration and control head update is described by a unitary U which is determined by the QTM transition rules. If the initial QTM configuration is $|\psi\rangle$, the state at the next time-step is $U|\psi\rangle$. A more detailed description is given in [BV97].

Similar to the PSPACE-hardness proof, we run TM2 and TM3 only if the computation outputs an accepting state. However, in the rejecting state, we choose to run a TM2', which is identical to TM2 except it does not flip $a_1 \to a_2$, but otherwise exactly mimics the motion of TM2. The analysis of the eigenvalues and eigenstates of the Hamiltonian proceeds identically: following Section 4.3.2. Since the encoded TMs are guaranteed to halt, we find that the eigenvalues of the effective Hamiltonian are given as $\lambda_j(H_{\text{eff}}) = 2\cos\left(\frac{2\pi j}{T}\right)$.

4.7.2 Time-Averaged Expectation Values

In order to see the long-time relaxation behaviour of the Hamiltonian, we need to observe the Hamiltonian for a sufficiently long time.

The only other difference is that the BQP computation will only accept with probability $> 2/3$ or $< 1/3$. Thus, we use a standard amplification procedure to push the acceptance probabilities to $1 - 2^{-\text{poly}\,N}$ and $2^{-\text{poly}\,N}$ respectively [MW05]. Using the same definition of $|w_k\rangle$ as

Section 4.3.2, we see that

$$\langle w_j|\mathcal{A}_N|w_j\rangle \begin{cases} = 0 & j < T_{halt}, \\ > (1-\alpha)(1-2^{-\operatorname{poly} N}) & j > T_{halt} + T_2 \text{ and TM1 accepts}, \\ < (1-\alpha)2^{-\operatorname{poly} N} & j > T_{halt} + T_2 \text{ and TM1 rejects}. \end{cases} \tag{4.92}$$

For the intermediate times, where TM2 is still acting, we note that in the rejecting case $\langle w_j|\mathcal{A}_N|w_j\rangle < (1-\alpha)2^{-\operatorname{poly} N}$.

To simplify the analysis, we first note that provided we are averaging over sufficiently long times, the averaged expectation value is close to the infinite time value. We prove the following in Appendix C.1:

Lemma 4.44. *Consider the time-averaged expectation value:*

$$\bar{\mathcal{A}}_N(\tau) = \frac{1}{\tau}\int_0^\tau dt \langle\psi_0|e^{iHt}\mathcal{A}_N e^{-iHt}|\psi_0\rangle. \tag{4.93}$$

Then, the following bound between the infinite time and finite time expectation values holds:

$$|\bar{\mathcal{A}}_N(\tau) - \bar{\mathcal{A}}_N| = O\left(\frac{T^2}{\tau}\right). \tag{4.94}$$

Hence we simply need to choose $\tau \gg T^2$ and our analysis will reduce to the analysis in infinite time limit. Since $T = \operatorname{poly} N$, for convenience we choose $\tau = p(N)T^2$, for some polynomial p, as an appropriate (but still polynomial) time period. Repeating the analysis of Section 4.3.3, we

find that

$$\bar{\mathcal{A}}_N(\tau) \begin{cases} > \frac{1}{2}(1-\alpha)(1-1/\text{poly}\, N) - O\left(\frac{1}{p(N)}\right) & \text{TM1 accepts,} \\ < O\left(\frac{1}{p(N)}\right) & \text{TM1 rejects.} \end{cases} \tag{4.95}$$

We can then verify that the promise is satisfied for this Hamiltonian provided in an identical way to Lemma 4.26, but noting that $T = O(\text{poly}\, N)$.

Theorem 4.45. FTFSRelax *is* BQP*-complete.*

Proof. By the above analysis, the value of $\bar{\mathcal{A}}_N(\tau)$ can be made to depend on whether an instance of a BQP-complete problem is accepted or rejected, so FTFSRelax is BQP-hard. Further, due to the promise, we can estimate $\bar{\mathcal{A}}_N$ to constant precision by simply simulating Hamiltonian evolution for polynomial time and measuring $\mathcal{A}_N$. Hence FTFSRelax $\in$ BQP, which implies FTFSRelax is BQP-complete. □

4.8 Discussion

We have shown there exist Hamiltonians for which, given an initial state and an observable to measure, predicting the long-time average of the observable is computationally intractable even for a quantum computer. An immediate consequence of this is that, assuming PSPACE does not collapse to a lower (easy) complexity class, there is no "easy" way of generically determining whether a given Hamiltonian and initial state thermalizes. This should be compared to the eigenstate thermalization hypothesis, which states that Hamiltonians satisfying certain conditions on their matrix elements should thermalize. Hence, either checking whether a Hamiltonian satisfies the conditions of the ETH on its matrix elements must not be computationally easy, or the ETH

must not be a full characterization of all Hamiltonians which thermalize.

We briefly discuss here the similarities and differences between our construction and that of Shiraishi and Matsumoto [SM21]. Primarily, their work focuses on systems in the thermodynamic limit (i.e. $N \to \infty$), while we focus on finite sized systems. While their result applies to almost-uniform product state inputs, our hardness construction is for non-uniform product state inputs in order to satisfy the promise of inverse-exponentially-separated eigenvalues for the eigenvectors with support on the input state. We take advantage of this relaxation to change the Turing machines used, as we do not need to decode the input to the URTM. We also make use of a more concise local Hilbert space for our hardness proof of FSRelax, although we currently require greater local dimension to prove the hardness of FSTherm(MC). Finally, as their result shows undecidability, they do not need to prove containment for their decision problem. Our contribution in this work also includes PSPACE algorithms to compute the long time observable average, as well as the microcanonical expectation values, and thereby decide FSTherm(MC).

There are several limitations to our work. In our tuning construction, we showed hardness for a restricted family of microcanonical states, with energy windows having inverse-logarithmic width ratio with the Hamiltonian norm. This is a technical restriction which arises due to the complexity of preparing the microcanonical state. In addition, in our proof of PSPACE-containment of FSTherm(MC), we assumed that $w = \Omega(\frac{1}{\text{poly } N})$. In future work, we hope to show PSPACE-completeness for the more general microcanonical state, with narrower energy windows. Our tuning construction also requires large local Hilbert space dimension. We hope to refine this result in future versions of this work. A hardness result with low local dimension would make our results much more applicable to 'natural' Hamiltonians.

Further Work While in Section 4.4 we showed that FSTherm(MC) is contained in PSPACE for both the Gibbs ensemble and the microcanonical ensemble with greater than inverse-polynomial width of the energy window, our hardness proof in Section 4.6 only applies to the microcanonical ensemble with an energy window $w = \Omega(\|H\|/\sqrt{\log N})$. Further, the hardness proof relies on a quantum polynomial time reduction as opposed to a more standard classical polynomial time reduction. In future work, we hope to show hardness under the more standard reduction, as well as for most general ensembles. In Section 4.6.4, we presented a conjecture under which FSTherm(Gibbs) is PSPACE-hard. If this conjecture is true, then combined with Lemma 4.36, FSTherm(Gibbs) is PSPACE-complete (and using only a classical polynomial time reduction). In future work, we hope to verify this conjecture. Our argument in Section 4.6.4 that there is a classical polynomial time reduction from FSHalt to FSTherm(Gibbs) (if our conjecture is true) makes use of the fact that in 1D for constant β, there exists a polynomial time classical algorithm to compute Gibbs expectation values [AC21]. This algorithms make use of the locality of H and exponential decay of correlations to construct matrix product operators for Gibbs states. In our tuning construction in Section 4.6.2 to prove hardness of FSTherm(MC), which relies on computing local observables, we do not make use of the additional structure of H when estimating local observables of ρ_{MC}. One possible strategy to make the reduction from FSHalt to FSTherm(MC) classical rather than quantum could be to make use of this structure to show that local observables of ρ_{MC} can be estimated efficiently classically. Such an approach may show that FSTherm(MC) is indeed PSPACE-complete.

One direction for future work is investigating complexity of different formulations of thermalization and equilibration. An example of a notion of thermalization and ensembles that is not captured by our work is equilibration of integrable systems to generalized Gibbs ensem-

bles [Rig+07; DAl+16]. An additional open question is related to thermalization of subsystems. For example, in a system with two parts A, B, with dynamics generated by a hamiltonian $H = H_A + H_B + H_{AB}$, we may like to know under which conditions and inverse temperatures β the behavior of subsystem A is captured by the Gibbs ensemble $\propto e^{-\beta H_A}$.

An alternative route to further research is to consider the reasons why Hamiltonians do not thermalize from a complexity-theoretic perspective. Some relevant thermalization-inhibiting phenomena include many-body localization (MBL) and quantum scars. While the complexity of simulating systems with MBL been studied previously [Hua15; Ehr+22], the authors of this work do not know of any results concerning the complexity of determining MBL from a description of the Hamiltonian.

Two additional obvious and interesting paths are as follows: finding constructions with smaller local Hilbert space dimensions, and proving a complexity result without the assumption on the gap between eigenvalues. Recent work has shown the presence of ultraslow relaxation in systems based on the complexity of word problems, with small local dimension [Bal+24]. In future work we hope to explore whether similar techniques can be used to yield results on the hardness of thermalization but with smaller local dimension.

Chapter 5: Outlook

Quantum information science elevates information to a physical quantity that can be studied much like energy and momentum. Information can be quantified, transferred, and treated as a resource in physical systems. In Chapter 1, we set out with the goal of investigating the dynamics of quantum information in local systems. These dynamics are important from two perspectives. The first is that of quantum computation. Scalable quantum architectures are constrained by locality and a natural way to implement interactions between their parts is by moving information around. By performing quantum routing, we may simulate all-to-all interactions, thereby overcoming connectivity constraints. However, implementing these permutations induces an overhead in circuit depth or the time taken for various quantum information tasks. It is hence crucial to find optimal protocols as well as lower bounds for this task.

The other perspective on information dynamics that we investigated arises from equilibration. The theory of statistical physics has been remarkably successful at describing the behavior of large, complex systems using a small number of macroscopic properties. One reason for this success is the fact that such complex systems evolve over sufficiently long times in such a way as to be well-approximated by statistical ensembles that only depend on the macroscopic properties of the initial state. The process by which systems evolve over time and observable averages take on fixed values is known as equilibration or relaxation. In the case of relaxing to the thermal state

(a statistical ensemble), this process is known as thermalization. While relaxation and thermalization in classical systems have been well understood, there has not been a conclusive answer elucidating their mechanism in quantum systems. This is due to the fact that unlike classical systems, isolated quantum systems evolve unitarily and hence cannot 'lose memory' of the initial state (and the large number of associate microscopic properties). Nevertheless, isolated quantum systems appear to thermalize and can be well-approximated by statistical ensembles. One proposed explanation for this has been due to information dynamics. It has been suggested that as quantum systems evolve, local information becomes spread and small subsystems become entangled with the rest of the system. This leads to local systems losing memory of the initial state and microscopic properties, which allows the system to be well-approximated by a statistical ensemble.

With these perspectives, our investigations aimed to address the following two questions:

Question 1. How fast can we engineer flow of information in quantum systems?

Question 2. How fast does information flow in quantum systems when we do not control it, and how hard is it to make predictions about such uncontrolled systems?

Chapter 2 and Chapter 3 of this thesis investigate the first question through the lens of quantum routing, while parts of Chapter 3 and Chapter 4 investigate the second through the lens of entanglement dynamics and thermalization. We briefly recap some of our main results. In Chapter 2, we investigated the use of mid-circuit measurements and classical feedback to speed up quantum routing. The ability to perform mid-circuit measurements and feedback enables the violation of Lieb-Robinson bounds on information transfer. We found classes of permutations and architecture graphs for which teleportation enabled a speedup over swap-based methods. We designed a graph for which teleportation obtains a $O(\log N)$ speedup over swap-based routing

even for worst-case permutations. Our main result was a bound on the advantage of teleportation routing over swap-based routing, showing that for worst-case permutations, on any graph the maximum speedup in routing time is $O(\sqrt{N \log N})$. In Chapter 3, we investigated quantum routing with continuous-time evolution in systems with a vertex bottleneck. We showed an improved lower bound on routing in such systems. We also designed an optimal routing protocol on the star graph for free particles, which obtains a polynomial speedup over gate-based routing.

Our investigation in Chapter 3 also addressed Question 2 by considering the case of *uncontrolled* flow of quantum information. We investigated the entanglement dynamics in systems with a vertex bottleneck. Existing results such as the small incremental entangling theorem (SIE) only consider the worst-case entanglement dynamics, which meant they could not provide a tight bound for routing. Using similar techniques to those developed for our routing lower bound, we showed a tighter upper bound on the amount of bipartite entanglement that can be generated in time t through a vertex bottleneck. Instead of just the worst-case, we considered the typical entanglement dynamics for random input states drawn from a 1-design.

In Chapter 4, we addressed Question 2 from the perspective of thermalization, making use of tools from computational complexity theory. While much research has focused on the question of when (and why) thermalization or relaxation occur, we show that this question is actually hard to decide even for a quantum computer. In particular, we showed that the problem of deciding whether a system thermalizes is PSPACE-complete.

At the end of each chapter, we have discussed corresponding conclusions and open questions in detail. Here, we present some discussions and directions for future investigation from the perspective of the full thesis. Chapter 3 indicates that the answer to Question 1 is subtle and depends on more than just the geometry of the architecture. Whether we have local ancillas, and the

strength of those local interactions, can greatly affect the time taken for routing. An example of this is routing on the vertex barbell. In the presence of ancillas (with arbitrarily fast interactions between the data qubits and their corresponding ancillas) we can perform routing much faster than our bound without ancillas, and we conjecture that it can be performed in $\mathrm{poly}(\log(N))$ time on the vertex barbell graph with N vertices. One possible direction for future research is investigating the possibility of speedups for quantum routing using both Hamiltonian evolution and LOCC. While we have constructed examples of speedups over gate-based routing for each case, we have not designed speedups that make use of both. Since Hamiltonian routing appears to have an advantage for vertex bottlenecks (which our teleportation routing model remains limited by), and teleportation routing has an advantage for permutations that make use of long paths (which Hamiltonian routing is limited by), it seems promising that a powerful routing model that makes use of both Hamiltonian routing and LOCC may obtain larger speedups.

Another possible direction for future work is investigating the complexity of decision problems related to entanglement dynamics. Whether or not a state is highly entangled has implications for its simulability and our ability to make predictions about its observables. The importance of entanglement as a resource and to shed light on quantum phenomena has prompted several investigations into identifying its presence, quantity, and structure in a variety of different scenarios [Gut+15; Ami+08; BZZ24; BGW24; Bak+24; BO20; Gha10]. While we showed that it is computationally complex to determine whether a system relaxes at long times, our results do not specifically address the entanglement of the state at long times. For general mixed states, the problem of deciding whether they are entangled or not (given as input the entries from their density matrix, which may themselves be hard to obtain) is known to be NP-hard [Gha10]. On the other hand, for pure states this problem is classically easy since pure state density matrices are

rank 1 (albeit still requiring the full classical description of the state). In a more restricted setting where only the Hamiltonian is given as input, it has been shown that learning the ground state entanglement is cryptographically hard [BZZ24; Bou+24]. Some related questions for future work are thus to determine the complexity of estimating the amount of entanglement, or the location of entanglement in thermal states. While Bakshi et al. [Bak+24] show that thermal entanglement vanishes at high temperatures, an open direction for investigation is thermal entanglement at low temperatures. Alternatively, one could investigate the same questions for states obtained after evolving a given initial under a given input Hamiltonian for a long time (and possibly equilibrating).

Finally, a potential direction for further related research is characterizing the time required for relaxation or thermalization to occur (or determining the computational complexity of classifying the thermalization timescale). Recently, there has been much interest in quantum algorithms to prepare Gibbs states and sample from them, particularly due to the classical computational hardness of this task [CKG23; BCL24]. Crucially, existing algorithms for Gibbs state preparation are only efficient in systems that have the property of mixing/thermalizing rapidly. One potential direction for future research is to further classify and identify properties of systems which mix rapidly and are hence amenable to such algorithms, as in Rouzé et al. [RFA24]. Alternatively, similarly to our result in Chapter 4, it may be possible to show that deciding whether a system mixes rapidly, or estimating the relaxation time are computationally complex.

Appendix A: Appendices for Chapter 2

A.1 Sparse routing

Previous work [Bap+23] shows an $O(\mathrm{diam}(G)+k^2)$ swap-based routing algorithm to route k vertices on a graph G. In this Appendix, we show that using swap s with a constant number of ancillas per qubit, this result can be improved to be linear in k.

We first introduce the following definitions.

Definition A.1.1 (Null token)**.** A *null token* is a dummy token that can be routed anywhere. In quantum routing, all ancillas are initialized with a null token in state $|0\rangle$.

Definition A.1.2 (Train)**.** A *train* is a set of non-null tokens along a path subgraph of G.

Now we show how a train can *advance*, i.e., translate by 1 along its length.

Lemma A.1.3. *A train can advance in depth 5.*

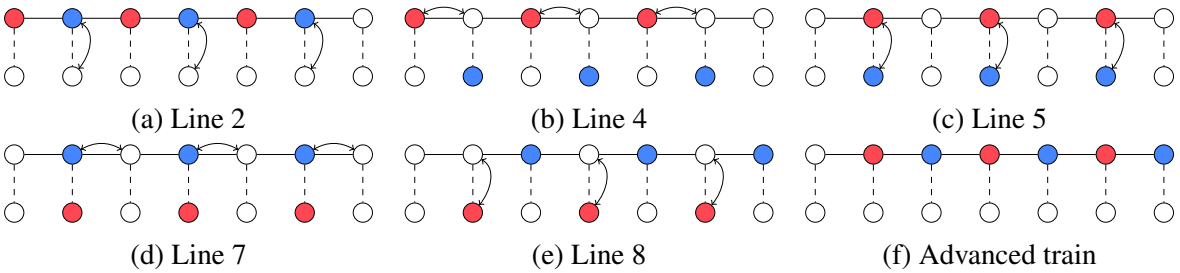
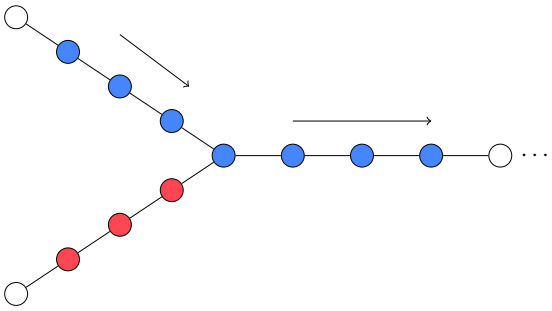
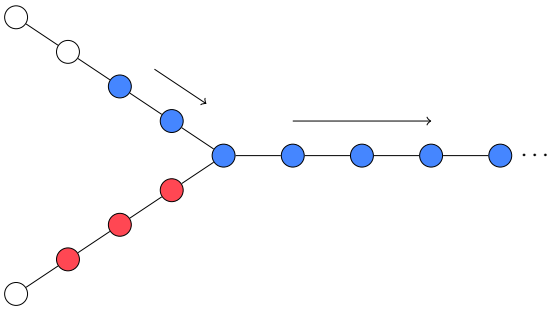


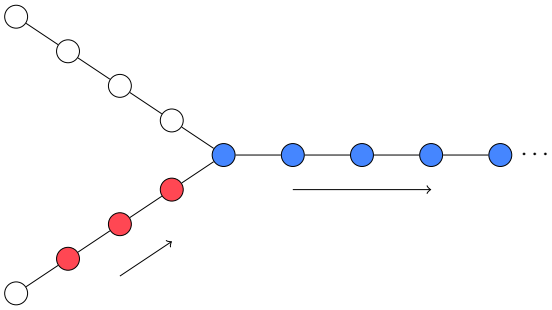
Figure A.1: Advancing a train of tokens, as in Algorithm A.1.1. Blank vertices hold state $|0\rangle$. The dashed lines represent connections with the local ancilla qubits.



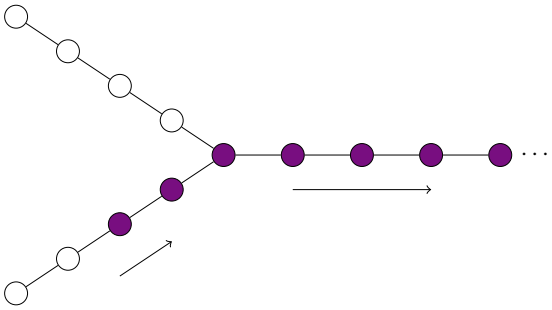
(a) The blue train advances one vertex at a time.



(b) The red train waits for the blue train, but remains in the same connected token cluster.



(c) The red train is concatenated to the blue train.



(d) Once concatenated, they move as a single train.

Figure A.2: Joining trains of tokens.

Algorithm A.1.1: Advance a train

Input: Train T from vertices 0 to $l-1$ on a path. Vertex l has a null token. The data token on vertex i is data(i), and the token on the corresponding ancilla is ancilla(i).

parallel for $i = 1, 3, \ldots$ **:**
  swap data(i) with ancilla(i)
parallel for $i = 0, 2, 4, \ldots$ **:**
  swap data(i) with data$(i+1)$
  swap data$(i+1)$ with ancilla$(i+1)$
parallel for $i = 1, 3, \ldots$ **:**
  swap data(i) with data$(i+1)$
  swap data(i) with ancilla(i)

Algorithm A.1.2: Token cluster movement

Input: Set of token clusters; vertex r.

1 In each token cluster, advance the train with head closest to r by 1 using Algorithm A.1.1.

2 If any two token clusters are adjacent, join them as a single token cluster.

3 If the head of any train T_1 is adjacent to the tail of another train T_2, join them as a single train with the head of T_2 and the tail of T_1.

Proof. Suppose we want to move a train of length l towards some vertex r. We define the *head* of a train as the token on the vertex closest to r, and the *tail* as the token on the vertex furthest from r. Consider the path subgraph spanned by the vertices the train lies on as well as the vertices of the shortest path from the head to r. Let the tail lie on vertex 0, and head lie at vertex $l-1$. We use Algorithm A.1.1 to advance the train such that after 5 time steps, the tail of the train is at vertex 1 and the head at l. This procedure is depicted in Fig. A.1. □

We now define a token cluster.

Definition A.1.4 (Token cluster)**.** A *token cluster* is a set of trains such that each train contains a token on a vertex that is adjacent to a vertex with a token from another train in the token cluster.

Token clusters move as in Fig. A.2, by Algorithm A.1.2. Once a train joins a token cluster, it remains connected and part of the token cluster.

We now prove Theorem 2.5, which we reproduce here for clarity.

Theorem 2.5 (Sparse routing)**.** *For any N-vertex connected simple graph G and $k \in [N]$,*

$$\mathrm{rt}_{\mathrm{swap}}(G,k) = O(\mathrm{diam}(G)+k). \tag{2.14}$$

Proof. Let us call the k tokens on vertices

$$\{v \in V(G) \mid \pi(v) \neq v\} \tag{A.1}$$

marked tokens, and let the remaining token be *unmarked tokens*. Our algorithm involves three phases.

Phase 1: First, we swap the unmarked tokens into local ancilla qubits and store them there for the duration of routing. Every vertex that initially held an unmarked token now holds a null token.

Next, we select some vertex r of G arbitrarily (in practice, selecting r to be at the center of the graph may provide constant-factor speedups). We now move all marked tokens towards vertex r by swap ping along the shortest possible paths, until the tokens span a set of vertices forming a tree connected to r. The tokens are moved in parallel, and when their paths intersect, the tokens move as trains, as per Lemma A.1.3. When the paths of multiple trains intersect, they form a token cluster, and can be moved as in Algorithm A.1.2.

Any given train is at most $\mathrm{diam}(G)$ distance away from r at the start of Phase 1. At every time step, a train either advances by 1 vertex towards r, or is part of a token cluster in which another train closer to r advances. Therefore, every token cluster becomes connected to r in depth $O(\mathrm{diam}(G))$, since in every token cluster, at least 1 train must reach r in depth $O(\mathrm{diam}(G))$. In particular, in $O(\mathrm{diam}(G))$ depth, all non-null tokens must span a tree containing r, and thus have merged into a single token cluster.

Phase 2: Now we have k vertices spanning a tree T. Suppose token v is mapped to the vertex $t(v)$ in T after Phase 1. Note that the token u that was originally at $t(v)$ must also be

a marked token, and therefore must now lie in T. We route the tokens on T according to a permutation π' such that

$$\pi'(t(v)) \coloneqq t(\pi(v)) \tag{A.2}$$

for all $t(v) \in V(T)$, in depth $2k$ [ACG94].

Phase 3: We now simply perform Phase 1 in reverse. During Phase 1, the marked token at u was mapped to $t(u)$. Therefore, after Phase 3, the token at $t(u)$ is mapped to vertex u. Therefore, the following mapping is applied to all vertices with marked tokens:

$$u \xrightarrow{\text{Phase 1}} t(u) \xrightarrow{\text{Phase 2}} t(\pi(u)) \xrightarrow{\text{Phase 3}} \pi(u). \tag{A.3}$$

The combined depth of the three phases is at most $O(k + \operatorname{diam}(G))$. □

A.2 Proof of diameter-expansion trade-off

In this appendix, we prove Lemma 2.7, adapting Proposition 3.1.5 from [Kow19] to vertex neighborhoods rather than edge neighborhoods.

Lemma 2.7. *For any connected simple graph G,*

$$\operatorname{diam}(G) \leq 2\frac{\log \frac{N}{2}}{\log\left(1 + c(G)\right)} + 2. \tag{2.28}$$

Proof. For any vertex $v \in V$, denote by $C(v, k)$ the set of all vertices that are at distance k from v. We call $C(v, k)$ a circle of radius k centered on v. Note that $C(v, k) \cap C(v, k') = \emptyset$ when

$k \neq k'$. Next, define

$$D(v,k) \coloneqq \bigcup_{r=0}^{k} C(v,r) \tag{A.4}$$

to be the disk of radius k centered on v. Observe that $C(v,0) = D(v,0) = \{v\}$. Finally, choose an integer $\rho(v)$ such that $|D(v,\rho(v))| \leq N/2 < |D(v,\rho(v)+1)|$ and call it the *horizon* of v. For any vertex, a horizon exists and is an integer between 0 and $\operatorname{diam}(G) - 1$.

By definition, for all $k \leq \rho(v) + 1$, we have

$$|C(v,k)| \geq c(G) \cdot |D(v,k-1)|. \tag{A.5}$$

Applying this inequality gives

$$|D(v,\rho(v))| = |C(v,\rho(v))| + |D(v,\rho(v)-1)| \tag{A.6}$$

$$\geq (1 + c(G))|D(v,\rho(v)-1)|. \tag{A.7}$$

Recursing until we reach the base case $D(v,0) = \{v\}$, we obtain

$$N/2 \geq (1 + c(G))^{\rho(v)}, \tag{A.8}$$

giving

$$\rho(v) \leq \frac{\log(N/2)}{\log(1 + c(G))}. \tag{A.9}$$

Next, for any two vertices $u, v \in V$, let $d(u,v)$ denote the distance between u, v. We claim that

$$d(u,v) \leq \rho(u) + \rho(v) + 2. \tag{A.10}$$

To see this, note that by definition, $|D(u, \rho(u) + 1)| > N/2$ and $|D(v, \rho(v) + 1)| > N/2$, which implies that $D(u, \rho(u) + 1) \bigcap D(v, \rho(v) + 1) \neq \emptyset$ by the pigeonhole principle. Therefore, there exists a vertex t such that $d(u, t) \leq \rho(u) + 1$ and $d(t, v) \leq \rho(v) + 1$. By the triangle inequality, we have $d(u, v) \leq \rho(u) + \rho(v) + 2$ as claimed.

Finally, we use Eq. (A.9) and maximize the distance over all vertex pairs u, v to get

$$\mathrm{diam}(G) \leq \frac{2 \log(N/2)}{\log(1 + c(G))} + 2 \tag{A.11}$$

as claimed. □

Appendix B: Appendices for Chapter 3

B.1 Proof of Lemma 3.12

In this section, we prove Lemma 3.12. First, we upper bound the Frobenius norms of commutators of terms in the Hamiltonian, which allows us to construct low-depth circuit approximations to the Hamiltonian evolution using the Trotter-Suzuki formula.

Lemma B.1.1. *Given a sequence $\eta_1, \ldots, \eta_{P+Q} \in \{LC, R\}$, let the nested commutator $[H_{\eta_{P+Q}}, \ldots, [H_{\eta_2}, H_{\eta_1}]]$ include H_{LC} a total of P times and H_R a total of Q times. Then there exists an $f(P,Q) > 0$ with no dependence on N_L, N_C, N_R such that*

$$\|[H_{\eta_{P+Q}}, \ldots, [H_{\eta_2}, H_{\eta_1}]]\|_{\mathrm{F}} \leq f(P,Q)\sqrt{N_L^P N_R^Q N_C}. \tag{B.1}$$

Proof. The proof strategy is as follows. We first show that the nested commutator can be expressed in terms of sums over Pauli strings with bounded coefficients, using induction to bound the coefficients. The base case is the first-order commutator. We write out terms of the commutator and regroup coefficients of identical Pauli strings that originate from products of different terms, and then bound the resulting coefficient. We take a similar approach to recursively identify the structure of the higher-order nested commutator and bound its coefficients. Finally, we use orthogonality of Pauli strings under the Frobenius norm to compute the bound of Eq. (B.1).

For convenience, we re-label the terms in H:

$$H^{(L)} = \sum_{l_1 \in L} h_{l_1}^{(L)\alpha_1} X_{l_1}^{\alpha_1}, \quad H^{(LL)} = \sum_{\substack{l_1, l_2 \in L \\ l_1 < l_2}} h_{l_1,l_2}^{(LL)\alpha_1\alpha_2} X_{l_1}^{\alpha_1} X_{l_2}^{\alpha_2}, \quad H^{(LC)} = \sum_{\substack{l_1 \in L \\ c_1 \in C}} h_{l_1,c_1}^{(LC)\alpha_1\gamma_1} X_{l_1}^{\alpha_1} X_{c_1}^{\gamma_1},$$

$$H^{(C)} = \sum_{c_1 \in C} h_{c_1}^{(C)\gamma_1} X_{c_1}^{\gamma_1}, \quad H^{(CC)} = \sum_{\substack{c_1, c_2 \in C \\ c_1 < c_2}} h_{c_1,c_2}^{(CC)\gamma_1\gamma_2} X_{c_1}^{\gamma_1} X_{c_2}^{\gamma_2},$$

$$H^{(R)} = \sum_{r_1 \in R} h_{r_1}^{(R)\beta_1} X_{r_1}^{\beta_1}, \quad H^{(RR)} = \sum_{\substack{r_1, r_2 \in R \\ r_1 < r_2}} h_{r_1,r_2}^{(RR)\beta_1\beta_2} X_{r_1}^{\beta_1} X_{r_2}^{\beta_2}, \quad H^{(RC)} = \sum_{\substack{r_1 \in R \\ c_1 \in C}} h_{r_1,c_1}^{(CR)\beta_1\gamma_1} X_{r_1}^{\beta_1} X_{c_1}^{\gamma_1} \tag{B.2}$$

(recall that we use the Einstein summation convention for $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2 \in \{1, 2, 3\}$). Here, $\forall\, i, l_i \in L, r_i \in R, c_i \in C$. We number the vertices in L, C, R uniquely so that these are disjoint sets. We group these terms as

$$H_{LC} = H^{(L)} + H^{(LL)} + H^{(LC)} + H^{(C)} + H^{(CC)}, \tag{B.3a}$$

$$H_R = H^{(R)} + H^{(RR)} + H^{(RC)}. \tag{B.3b}$$

As mentioned above, we proceed by induction. For any sequence $\eta_1, \ldots, \eta_{P+Q} \in \{LC, R\}$, suppose there are $P \geq 1$ LCs and $Q \geq 1$ Rs (the order does not matter). We will show by induction that the corresponding nested commutator, which we call a (P, Q)-commutator, is of

the form

$$[H_{\eta_{P+Q}},\ldots,[H_{\eta_2},H_{\eta_1}]] = \sum_{p_1=0}^{P}\sum_{p_2=1}^{P+1-p_1}\sum_{q=0}^{Q}\sum_{l_1<l_2<\cdots<l_{p_1}}\sum_{c_1<c_2<\cdots<c_{p_2}}\sum_{r_1<r_2<\cdots<r_q} A^{\alpha_1\cdots\alpha_{p_1},\beta_1\cdots\beta_q,\gamma_1\cdots\gamma_{p_2}}_{l_1\cdots l_{p_1},r_1\cdots r_q,c_1\cdots c_{p_2};P,Q} X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1}\cdots X^{\gamma_{p_2}}_{c_{p_2}}, \tag{B.4}$$

where

$$|A^{\alpha_1\cdots\alpha_{p_1},\beta_1\cdots\beta_q,\gamma_1\cdots\gamma_{p_2}}_{l_1\cdots l_{p_1},r_1\cdots r_q,c_1\cdots c_{p_2};P,Q}| \leq c_{P,Q} {N_L}^{(P-p_1)/2} {N_R}^{(Q-q)/2} {N_C}^{(1-p_2)/2}, \tag{B.5}$$

and where $c_{P,Q}$ is a constant that does not depend on N_L, N_C, N_R.

The base case is the first-order commutator

$$[H_{LC},H_R] = \sum_{c_1,r_1,c_1'} h^{(C)\gamma_1}_{c_1} h^{(CR)\beta_1\gamma_1'}_{r_1,c_1'} X^{\beta_1}_{r_1}[X^{\gamma_1}_{c_1},X^{\gamma_1'}_{c_1'}] + \sum_{c_1,c_2,r_1,c_1'} h^{(CC)\gamma_1\gamma_2}_{c_1,c_2} h^{(CR)\beta_1\gamma_1'}_{r_1,c_1'} X^{\beta_1}_{r_1}[X^{\gamma_1}_{c_1}X^{\gamma_2}_{c_2},X^{\gamma_1'}_{c_1'}]$$
$$+ \sum_{l_1,c_1,r_1,c_1'} h^{(LC)\alpha_1\gamma_1}_{l_1,c_1} h^{(CR)\beta_1\gamma_1'}_{r_1,c_1'} X^{\beta_1}_{r_1} X^{\alpha_1}_{l_1}[X^{\gamma_1}_{c_1},X^{\gamma_1'}_{c_1'}]. \tag{B.6}$$

By expanding the commutators and grouping terms based on which sites they are supported (i.e., act non-trivially) on, we can rewrite this in the form

$$[H_{LC},H_R] = \sum_{r_1,c_1} A^{\beta_1\tilde{\gamma_1}}_{r_1c_1} X^{\beta_1}_{r_1} X^{\tilde{\gamma_1}}_{c_1} + \sum_{r_1,c_1,c_2} A^{\beta_1\tilde{\gamma_1}\tilde{\gamma_2}}_{r_1c_1c_2} X^{\beta_1}_{r_1} X^{\tilde{\gamma_1}}_{c_1} X^{\tilde{\gamma_2}}_{c_2} + \sum_{l_1,r_1,c_1} A^{\alpha_1\beta_1\tilde{\gamma_1}}_{l_1r_1c_1} X^{\alpha_1}_{l_1} X^{\beta_1}_{r_1} X^{\tilde{\gamma_1}}_{c_1}. \tag{B.7}$$

$A^{\beta_1\tilde{\gamma_1}}_{r_1c_1}, A^{\alpha_1\beta_1\tilde{\gamma_1}}_{l_1r_1c_1}, A^{\beta_1\tilde{\gamma_1}\tilde{\gamma_2}}_{r_1c_1c_2}$ are coefficients of distinct Pauli strings in Eq. (B.7). For conciseness, we leave P, Q implicit in our notation for the first-order commutator, where $P = Q = 1$. Also note that, for example, the coefficients $A^{\alpha_1\beta_1\tilde{\gamma_1}}_{l_1r_1c_1}$ and $A^{\beta_1\tilde{\gamma_1}\tilde{\gamma_2}}_{r_1c_1c_2}$ are distinguishable since the lower indices

come from distinct sets (e.g., $l_1 \in L$ and $r_1 \in R$). Each $A^{\beta_1\tilde{\gamma_1}}_{r_1c_1}$ is a sum of coefficients of $X^{\beta}_{r_1}X^{\tilde{\gamma_1}}_{c_1}$ from Eq. (B.6):

$$A^{\beta_1\tilde{\gamma_1}}_{r_1c_1} = 2\sum_{\gamma_1,\gamma_1'} h^{(C)\gamma_1}_{c_1} h^{(CR)\beta_1\gamma_1'}_{r_1,c_1} \mathbb{I}[X^{\gamma_1}_{c_1}X^{\gamma_1'}_{c_1} = X^{\tilde{\gamma_1}}_{c_1}], \tag{B.8}$$

where $\mathbb{I}[x]$ is the indicator function that returns 0 (1) if x is false (true). For each $A^{\beta_1\tilde{\gamma_1}}_{r_1c_1}$, there are four contributing terms from Eq. (B.6) (counting the multiplicity factor 2 in Eq. (B.8)), obtained by expanding the commutator and making use of the Pauli multiplication rules. The number of such terms is the number of valid indices $(c_1, r_1, c_1', \beta_1, \gamma_1, \gamma_1')$ that, on regrouping Eq. (B.6) to the form Eq. (B.7), contribute to terms with new indices $(c_1, r_1, \beta_1, \tilde{\gamma_1})$. A tilde on an index indicates that it is an updated index in the commutator formed by the multiplication of two Paulis as in Eq. (B.8), where $\tilde{\gamma_1}$ is the index corresponding to the single-qubit Pauli obtained on multiplying X^{γ_1} and $X^{\gamma_1'}$. Indices without a tilde are unchanged and are identical to those in H_{LC}, H_R.

Likewise, each $A^{\alpha_1\beta_1\tilde{\gamma_1}}_{l_1r_1c_1}$ is the sum of four coefficients from Eq. (B.6), and each $A^{\beta_1\tilde{\gamma_1}\tilde{\gamma_2}}_{r_1c_1c_2}$ is the sum of eight coefficients.

Hence

$$|A^{\beta_1\tilde{\gamma_1}}_{r_1c_1}| \leq 4\max|h^{(C)\gamma_1}_{c_1} h^{(CR)\beta_1\gamma_1}_{r_1,c_1}| \leq 4\sqrt{N}, \tag{B.9a}$$

$$|A^{\alpha_1\beta_1\tilde{\gamma_1}}_{l_1r_1c_1}| \leq 4\max|h^{(LC)\alpha_1\gamma_1}_{l_1,c_1} h^{(CR)\beta_1\gamma_1}_{r_1,c_1}| \leq 4, \tag{B.9b}$$

$$|A^{\beta_1\tilde{\gamma_1}\tilde{\gamma_2}}_{r_1c_1c_2}| \leq 8\max|h^{(CC)\gamma_1\gamma_2}_{c_1,c_2} h^{(CR)\beta_1\gamma_1'}_{r_1,c_1'}| \leq 8. \tag{B.9c}$$

Now, as different Pauli strings are orthogonal with respect to the Frobenius norm, we have

$$\begin{aligned}\|[H_{LC}, H_R]\|_{\mathrm{F}}^2 &\leq \sum_{c_1,r_1} 16N + \sum_{c_1,c_2,r_1} 64 + \sum_{l_1,c_1,r_1} 16 \\ &\leq 16N \times N_C \times N_R + 64N_C^2 \times N_R + 16N_L \times N_C \times N_R, \end{aligned} \tag{B.10}$$

satisfying Eq. (B.1) for $P = Q = 1$ because we have assumed N_L is larger than N_R, N_C.

We now consider higher-order commutators. We use the induction hypothesis Eq. (B.4) to prove that the same equation holds for the $(P+1, Q)$-commutator, i.e., adding one more outermost commutator with H_{LC}. The case $\widetilde{P} = P$, $\widetilde{Q} = Q + 1$ follows analogously. We use a tilde to denote next-order quantities, e.g., the numbers of LCs and Rs in the new η sequence are

$$\widetilde{P} = P + 1, \qquad \widetilde{Q} = Q. \tag{B.11}$$

We use $\cdots$ to indicate products over intermediate indices. For example, $X_{l_1}^{\alpha_1} \cdots X_{l_j}^{\alpha_j}$ indicates $\prod_{i=1}^{j} X_{l_i}^{\alpha_i}$. We also use X_j^0 to denote the identity operator for any j. We write $l \notin \overrightarrow{l_{p_1}}$ as shorthand for $l \notin \{l_1, \ldots, l_{p_1}\}$, and similarly for $c \notin \overrightarrow{c_{p_2}}$, etc. We also abbreviate $A_{l_1 \cdots l_{p_1}, r_1 \cdots r_q, c_1 \cdots c_{p_2}; P, Q}^{\alpha_1 \cdots \alpha_{p_1}, \beta_1 \cdots \beta_q, \gamma_1 \cdots \gamma_{p_2}}$ as $A_{\overrightarrow{l_{p_1}}, \overrightarrow{r_q}, \overrightarrow{c_{p_2}}; P, Q}^{\overrightarrow{\alpha_{p_1}}, \overrightarrow{\beta_q}, \overrightarrow{\gamma_{p_2}}}$. We first compute the commutators of terms in H_{LC} (whose indices are labeled by primes, e.g., $X_{l_1'}^{\alpha_1'} X_{c_1'}^{\gamma_1'}$) with terms from the (P, Q)-commutator (whose indices are labeled without primes, e.g., $X_{l_1}^{\alpha_1} \cdots X_{l_{p_1}}^{\alpha_{p_1}} X_{r_1}^{\beta_1} \cdots X_{r_q}^{\beta_q} X_{c_1}^{\gamma_1} \cdots X_{c_{p_2}}^{\gamma_{p_2}}$).

First, we consider the commutator with $H^{(L)}$. Any term $X_{l_1'}^{\alpha_1'}$ in $H^{(L)}$ is non-commuting

with $X_{l_1}^{\alpha_1}\cdots X_{l_{p_1}}^{\alpha_{p_1}}X_{r_1}^{\beta_1}\cdots X_{r_q}^{\beta_q}X_{c_1}^{\gamma_1}\cdots X_{c_{p_2}}^{\gamma_{p_2}}$ iff $\exists\, j$ such that $l_j = l'_1$ and $\alpha_j \neq \alpha'_1$. Therefore,

$$
\begin{aligned}
&[H^{(L)}, A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} X_{l_1}^{\alpha_1}\cdots X_{l_{p_1}}^{\alpha_{p_1}}X_{r_1}^{\beta_1}\cdots X_{r_q}^{\beta_q}X_{c_1}^{\gamma_1}\cdots X_{c_{p_2}}^{\gamma_{p_2}}] \\
&\quad = \sum_{l'_1\in L} h_{l'_1}^{(L)\alpha'_1} A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} \mathbb{I}[l'_1 = l_j] X_{l_1}^{\alpha_1}\cdots X_{l_j}^{0}\cdots X_{l_{p_1}}^{\alpha_{p_1}}X_{r_1}^{\beta_1}\cdots X_{r_q}^{\beta_q}X_{c_1}^{\gamma_1}\cdots X_{c_{p_2}}^{\gamma_{p_2}}[X_{l'_1}^{\alpha'_1}, X_{l_j}^{\alpha_j}].
\end{aligned}
\tag{B.12}
$$

Now we consider terms in $H^{(LL)}$. For a term $X_{l'_1}^{\alpha'_1}X_{l'_2}^{\alpha'_2}$, we divide the terms from the (P,Q)-commutator which have the form $X_{l_1}^{\alpha_1}\cdots X_{l_{p_1}}^{\alpha_{p_1}}X_{r_1}^{\beta_1}\cdots X_{r_q}^{\beta_q}X_{c_1}^{\gamma_1}\cdots X_{c_{p_2}}^{\gamma_{p_2}}$ into those with exactly one overlapping site with l'_1, l'_2 or exactly two overlapping sites. The commutator is thus expressed as

$$
\begin{aligned}
&[H^{(LL)}, A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} X_{l_1}^{\alpha_1}\cdots X_{l_{p_1}}^{\alpha_{p_1}}X_{r_1}^{\beta_1}\cdots X_{r_q}^{\beta_q}X_{c_1}^{\gamma_1}\cdots X_{c_{p_2}}^{\gamma_{p_2}}] \\
&\quad = \sum_{\substack{l_1,l_2\in L\\ l_1<l_2}} h_{l'_1,l'_2}^{(LL)\alpha'_1\alpha'_2} A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} X_{r_1}^{\beta_1}\cdots X_{r_q}^{\beta_q}X_{c_1}^{\gamma_1}\cdots X_{c_{p_2}}^{\gamma_{p_2}}\times \\
&\qquad \Bigg(\mathbb{I}[l'_1 = l_j]\mathbb{I}[l'_2 \notin \overrightarrow{l_{p_1}}] X_{l_1}^{\alpha_1}\cdots X_{l_j}^{0}\cdots X_{l_{p_1}}^{\alpha_{p_1}}X_{l'_2}^{\alpha'_2}[X_{l'_1}^{\alpha'_1}, X_{l_j}^{\alpha_j}] \qquad \text{(B.13)} \\
&\qquad + \mathbb{I}[l'_1 \notin \overrightarrow{l_{p_1}}]\mathbb{I}[l'_2 = l_j] X_{l_1}^{\alpha_1}\cdots X_{l_j}^{0}\cdots X_{l_{p_1}}^{\alpha_{p_1}}X_{l'_1}^{\alpha'_1}[X_{l'_2}^{\alpha'_2}, X_{l_j}^{\alpha_j}] \qquad \text{(B.14)} \\
&\qquad + \mathbb{I}[l'_1 = l_{j_1}]\mathbb{I}[l'_2 = l_{j_2}] X_{l_1}^{\alpha_1}\cdots X_{l_{j_1}}^{0}\cdots X_{l_{j_2}}^{0}\cdots X_{l_{p_1}}^{\alpha_{p_1}}[X_{l'_1}^{\alpha'_1}X_{l'_2}^{\alpha'_2}, X_{l_{j_1}}^{\alpha_{j_1}}X_{l_{j_2}}^{\alpha_{j_2}}] \Bigg). \qquad \text{(B.15)}
\end{aligned}
$$

Now we look at the commutator with $H^{(LC)}$. Similarly to $H^{(LL)}$, we divide terms of the order-$(P+Q)$ nested commutator into those that have exactly one or two overlapping sites with each

term of $H^{(LC)}$. Therefore,

$$
\begin{aligned}
&[H^{(LC)}, A^{\overrightarrow{\alpha_{p1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p2}};P,Q} X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1}\cdots X^{\gamma_{p_2}}_{c_{p_2}}] \\
&= \sum_{\substack{l_1\in L\\ c_1\in C}} h^{(LC)\alpha_1'\gamma_1'}_{l_1',c_1'} A^{\overrightarrow{\alpha_{p1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p2}};P,Q} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q}\times \\
&\quad \Bigg(\mathbb{I}[l_1'=l_j]\mathbb{I}[c_1'=c_k] X^{\alpha_1}_{l_1}\cdots X^{0}_{l_j}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\gamma_1}_{c_1}\cdots X^{0}_{c_k}\cdots X^{\gamma_{p_2}}_{c_{p_2}} [X^{\alpha_1'}_{l_1'} X^{\gamma_1'}_{c_1'}, X^{\alpha_j}_{l_j} X^{\gamma_k}_{c_k}] \qquad \text{(B.16)} \\
&\qquad + \mathbb{I}[l_1'\notin \overrightarrow{l_{p_1}}]\mathbb{I}[c_1'=c_j] X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\alpha_1'}_{l_1'} X^{\gamma_1}_{c_1}\cdots X^{0}_{c_j}\cdots X^{\gamma_{p_2}}_{c_{p_2}} [X^{\gamma_1'}_{c_1'}, X^{\gamma_j}_{c_j}] \qquad \text{(B.17)} \\
&\qquad + \mathbb{I}[l_1'=l_j]\mathbb{I}[c_1'\notin \overrightarrow{c_{p2}}] X^{\alpha_1}_{l_1}\cdots X^{0}_{l_j}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\gamma_1}_{c_1}\cdots X^{\gamma_{p_2}}_{c_{p_2}} X^{\gamma_1'}_{c_1'} [X^{\alpha_1'}_{l_1'}, X^{\alpha_j}_{l_j}] \Bigg). \qquad \text{(B.18)}
\end{aligned}
$$

Now we consider the commutator with $H^{(C)}$, which is analogous to the commutator with $H^{(L)}$:

$$
\begin{aligned}
&[H^{(C)}, A^{\overrightarrow{\alpha_{p1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p2}};P,Q} X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1}\cdots X^{\gamma_{p_2}}_{c_{p_2}}] \\
&\quad = \sum_{c_1'\in C} h^{(C)\gamma_1'}_{c_1'} A^{\overrightarrow{\alpha_{p1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p2}};P,Q} \mathbb{I}[c_1'=c_j] X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1}\cdots X^{0}_{c_j}\cdots X^{\gamma_{p_2}}_{c_{p_2}} [X^{\gamma_1'}_{c_1'}, X^{\gamma_j}_{c_j}].
\end{aligned}
\qquad \text{(B.19)}
$$

The case for $H^{(CC)}$ is analogous to the commutator with $H^{(LL)}$:

$$
\begin{aligned}
&[H^{(CC)}, A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1}\cdots X^{\gamma_{p_2}}_{c_{p_2}}] \\
&\quad = \sum_{\substack{c_1,c_2\in C\\ c_1<c_2}} h^{(CC)\gamma_1'\gamma_2'}_{c_1',c_2'} A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q}\times \\
&\qquad \Bigg(\mathbb{I}[c_1'=c_j]\mathbb{I}[c_2'\notin\overrightarrow{c_{p_2}}] X^{\gamma_1}_{c_1}\cdots X^{0}_{c_j}\cdots X^{\gamma_{p_2}}_{c_{p_2}} X^{\gamma_2'}_{c_2'}[X^{\gamma_1'}_{c_1'}, X^{\gamma_j}_{c_j}] \qquad \text{(B.20)} \\
&\qquad + \mathbb{I}[c_1'\notin\overrightarrow{c_{p_2}}]\mathbb{I}[c_2'=c_j] X^{\gamma_1}_{c_1}\cdots X^{0}_{c_j}\cdots X^{\gamma_{p_2}}_{c_{p_2}} X^{\gamma_1'}_{c_1'}[X^{\gamma_2'}_{c_2'}, X^{\gamma_j}_{c_j}] \qquad \text{(B.21)} \\
&\qquad + \mathbb{I}[c_1'=c_{j_1}]\mathbb{I}[c_2'=c_{j_2}] X^{\gamma_1}_{c_1}\cdots X^{0}_{c_{j_1}}\cdots X^{0}_{c_{j_2}}\cdots X^{\gamma_{p_2}}_{c_{p_2}}[X^{\gamma_1'}_{c_1'}X^{\gamma_2'}_{c_2'}, X^{\gamma_{j_1}}_{c_{j_1}}X^{\gamma_{j_2}}_{c_{j_2}}] \Bigg). \qquad \text{(B.22)}
\end{aligned}
$$

Overall, the $(P+1,Q)$-commutator is

$$
\begin{aligned}
[H_{LC},[H_{\eta_{P+Q}},\cdots,[H_{\eta_2},H_{\eta_1}]]] &= \sum_{p_1=0}^{P}\sum_{p_2=1}^{P+1-p_1}\sum_{q=0}^{Q}\sum_{l_1<l_2<\cdots<l_{p_1}}\sum_{c_1<c_2<\cdots<c_{p_2}}\sum_{r_1<r_2<\cdots<r_q}\Bigg(A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q} \\
&\times [H^{(L)}+H^{(LL)}+H^{(LC)}+H^{(C)}+H^{(CC)}, X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1}\cdots X^{\gamma_{p_2}}_{c_{p_2}}]\Bigg).
\end{aligned} \qquad \text{(B.23)}
$$

Using Eqs. (B.12) to (B.22) and grouping terms together, we can rewrite this as

$$
\begin{aligned}
&[H_{LC},[H_{\eta_{P+Q}},\cdots,[H_{\eta_2},H_{\eta_1}]]] \\
&= \sum_{\widetilde{p_1}=0}^{P+1}\sum_{\widetilde{p_2}=1}^{P+2-\widetilde{p_1}}\sum_{q=0}^{Q}\sum_{\widetilde{l_1}<\cdots<\widetilde{l_{\widetilde{p_1}}}}\sum_{\widetilde{c_1}<\cdots<\widetilde{c_{\widetilde{p_2}}}}\sum_{r_1<r_2<\cdots<r_q} A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}\times \\
&\qquad\qquad X^{\widetilde{\alpha_1}}_{\widetilde{l_1}}\cdots X^{\widetilde{\alpha_{\widetilde{p_1}}}}_{l_{\widetilde{p_1}}} X^{\widetilde{\beta_1}}_{r_1}\cdots X^{\widetilde{\beta_q}}_{r_q} X^{\widetilde{\gamma_1}}_{\widetilde{c_1}}\cdots X^{\widetilde{\gamma_{\widetilde{p_2}}}}_{c_{\widetilde{p_2}}}.
\end{aligned} \qquad \text{(B.24)}
$$

Each coefficient $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ is a sum of coefficients of the

form $A^{\overrightarrow{\alpha_{p_1}},\vec{\beta}_q,\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\vec{r}_q,\overrightarrow{c_{p_2}};P,Q} \times h$, where h is a coefficient from H_{LC}. We can thus bound $\|A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}\|$ using the induction hypothesis, which bounds $\|A^{\alpha_1\cdots\alpha_{p_1},\beta_1\cdots\beta_q,\gamma_1\cdots\gamma_{p_2}}_{l_1\cdots l_{p_1},r_1\cdots r_q,c_1\cdots c_{p_2};P,Q}\|$, and by counting the number of terms (and their norms) in each of Eqs. (B.12) to (B.22) that contribute to the $(P+1,Q)$ commutator:

- In Eq. (B.12), the number of sites in the support of the commutator is the same as the number of sites in the support of $X^{\alpha_1}_{l_1}\cdots X^{\alpha_{p_1}}_{l_{p_1}} X^{\beta_1}_{r_1}\cdots X^{\beta_q}_{r_q} X^{\gamma_1}_{c_1}\cdots X^{\gamma_{p_2}}_{c_{p_2}}$, i.e., $\widetilde{p_1}=p_1, \widetilde{p_2}=p_2$. We count the number of terms in the $(P+1,Q)$ commutator that are obtained from Eq. (B.12). For a given $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$, we count the number of terms from Eq. (B.12) with $l_1', \overrightarrow{l_{p_1}}$ that would produce a term in the $(P+1,Q)$-commutator with $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$. We see that $\overrightarrow{l_{p_1}}$ must equal $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$, and there are $\widetilde{p_1}$ choices of l_1' that overlap with a site from $\overrightarrow{l_{p_1}}$. As in Eq. (B.9a), we have a multiplicity of 4, obtained by expanding the commutator and from the Pauli multiplication rules. Thus, there are at most $4\widetilde{p_1}$ terms from Eq. (B.12) which are grouped together in Eq. (B.24) for each coefficient $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$.
- Similar to the analysis of Eq. (B.12), we count the number of terms of Eq. (B.13) that contribute to a given $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$. In Eq. (B.13), the number of sites in L in the support of the commutator increases by 1, so we have $\widetilde{p_1}=p_1+1, \widetilde{p_2}=p_2$. There are $\widetilde{p_1}$ choices of l_2' and $\widetilde{p_1}-1$ choices of l_1' that match a given $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$. Hence there are at most $4\widetilde{p_1}(\widetilde{p_1}-1)$ coefficients from Eq. (B.23) that are grouped together in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$.
- Eq. (B.14) is similar to Eq. (B.13). There are at most $4\widetilde{p_1}(\widetilde{p_1}-1)$ coefficients from Eq. (B.23) that are grouped together in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$.
- Eq. (B.15) is non-zero on terms of $H^{(LL)}$ which have two overlapping sites with terms

of the (P,Q)-commutator, such that the Paulis commute on one site and anticommute on the other site. An example of this on two sites is $[X_1X_2, X_1Z_2] \propto Y_2$. Once again, we count the number of choices of sites $l_1', l_2', \overrightarrow{l_{p_1}}$ in Eq. (B.15) that contribute to a given $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$, i.e., match the sites $\overrightarrow{\widetilde{l_{\widetilde{p_1}}}}$. There are at most N_L choices of the site to be removed (l_2'), and $\widetilde{p_1}$ choices of the site to retain (l_1'). For a given Pauli on site l_1', there are 6 commutators that contribute to the coefficient. For example, Y_2 can be obtained from the following commutators: $[X_1X_2, X_1Z_2], [Y_1X_2, Y_1Z_2], [Z_1X_2, Z_1Z_2]$ and the commutators obtained from these by swapping $X_2 \leftrightarrow Z_2$. On expanding each commutator, we obtain two terms. There are thus at most $12N_L\widetilde{p_1}$ contributing terms from Eq. (B.15) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1 - 1, \widetilde{p_2} = p_2$.

- Eq. (B.16) is similar to Eq. (B.15). Terms in Eq. (B.16) have one overlapping site in C and one overlapping site in L with the (P,Q)-commutator. Similar to Eq. (B.15), the number of sites decreases, in either C or L. If the site removed is from C, then there are at most N_C choices of the site to be removed (c_1'), and $\widetilde{p_1}$ choices of the site to retain (l_1'). Likewise, if the site removed is from L, then there are N_L choices of l_1' and $\widetilde{p_2}$ choices of c_1'. There are at most $12N_L\widetilde{p_2}$ terms of Eq. (B.16) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1 - 1, \widetilde{p_2} = p_2$. Likewise, there are at most $12N_C\widetilde{p_1}$ terms of Eq. (B.16) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1, \widetilde{p_2} = p_2 - 1$.
- From Eq. (B.17), the number of sites in L in the support of the commutator increases by one, while the number of sites in the support from C remains unchanged. Therefore, there are at most $4\widetilde{p_1}\widetilde{p_2}$ terms of Eq. (B.17) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1} = p_1 + 1, \widetilde{p_2} = p_2$.

- Eq. (B.18) is similar to Eq. (B.17), increasing the number of sites from C while leaving the number of sites from L unchanged. There are at most $4\widetilde{p_1}\widetilde{p_2}$ terms of Eq. (B.17) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{p_1}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{p_2}}}_{\widetilde{l_1}\cdots\widetilde{l_{p_1}},\widetilde{c_1}\cdots\widetilde{c_{p_2}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1}=p_1,\widetilde{p_2}=p_2+1$.

- Eq. (B.19) is similar to Eq. (B.12). There are at most $4\widetilde{p_2}$ terms of Eq. (B.19) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{p_1}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{p_2}}}_{\widetilde{l_1}\cdots\widetilde{l_{p_1}},\widetilde{c_1}\cdots\widetilde{c_{p_2}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1}=p_1,\widetilde{p_2}=p_2$.

- The analysis of Eq. (B.20) is similar to that of Eq. (B.13). There are at most $4\widetilde{p_2}(\widetilde{p_2}-1)$ terms of Eq. (B.13) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{p_1}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{p_2}}}_{\widetilde{l_1}\cdots\widetilde{l_{p_1}},\widetilde{c_1}\cdots\widetilde{c_{p_2}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1}=p_1,\widetilde{p_2}=p_2+1$.

- The analysis of Eq. (B.21) is similar to that of Eq. (B.20). There are at most $4\widetilde{p_2}(\widetilde{p_2}-1)$ terms of Eq. (B.13) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{p_1}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{p_2}}}_{\widetilde{l_1}\cdots\widetilde{l_{p_1}},\widetilde{c_1}\cdots\widetilde{c_{p_2}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_1}=p_1,\widetilde{p_2}=p_2+1$.

- The analysis of Eq. (B.22) is similar to that of Eq. (B.15). Eq. (B.22) is non-zero on terms of $H^{(CC)}$ which have two overlapping sites with terms of the (P,Q)-commutator, such that the Paulis on one site anticommute, and on the other site they commute. We count the number of choices of sites $c_1', c_2', \overrightarrow{c_{p_2}}$ that match with a given $\overrightarrow{\widetilde{c_{\widetilde{p_2}}}}$. There are at most N_C choices of the site to remove (c_2') and $\widetilde{p_2}$ choices of the site to retain (c_1'). Hence there are at most $12N_C\widetilde{p_2}$ terms of Eq. (B.22) grouped in $A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{p_1}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{p_2}}}_{\widetilde{l_1}\cdots\widetilde{l_{p_1}},\widetilde{c_1}\cdots\widetilde{c_{p_2}},r_1\cdots r_q;P+1,Q}$ with $\widetilde{p_2}=p_2-1,\widetilde{p_1}=p_1$.

Eq. (B.24) is the sum of Eqs. (B.12) to (B.22). Applying the triangle inequality to the cases above, we thus have

$$\begin{aligned}\|A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{p_1}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{p_2}}}_{\widetilde{l_1}\cdots\widetilde{l_{p_1}},\widetilde{c_1}\cdots\widetilde{c_{p_2}},r_1\cdots r_q;P+1,Q}\| &\leq 4\widetilde{p_1}\max\|A^{\overrightarrow{\alpha_{p_1}},\overrightarrow{\beta_q},\overrightarrow{\gamma_{p_2}}}_{\overrightarrow{l_{p_1}},\overrightarrow{r_q},\overrightarrow{c_{p_2}};P,Q}\|\Big|_{p_1=\widetilde{p_1},p_2=\widetilde{p_2}}+\cdots\\ &\leq 4\widetilde{p_1}\times c_{P,Q}{N_L}^{(P-\widetilde{p_1})/2}{N_R}^{(Q-q)/2}{N_C}^{(1-\widetilde{p_2})/2}+\cdots, \qquad \text{(B.25)}\end{aligned}$$

where the first term comes from Eq. (B.12), and $\cdots$ represents the contributions from the remaining equations. The second line follows from the induction hypothesis [Eq. (B.5)]. Adding the terms from the analysis of each of Eqs. (B.12) to (B.22), we thus have

$$
\begin{aligned}
&\|A^{\widetilde{\alpha_1}\cdots\widetilde{\alpha_{\widetilde{p_1}}},\widetilde{\beta_1}\cdots\widetilde{\beta_q},\widetilde{\gamma_1}\cdots\widetilde{\gamma_{\widetilde{p_2}}}}_{\widetilde{l_1}\cdots\widetilde{l_{\widetilde{p_1}}},\widetilde{c_1}\cdots\widetilde{c_{\widetilde{p_2}}},r_1\cdots r_q;P+1,Q}\| \\
&\leq \Bigg[(4\widetilde{p_1}+4\widetilde{p_2}) + \left(8\widetilde{p_1}^2+4\widetilde{p_1}\widetilde{p_2}-8\widetilde{p_1}\right)\sqrt{N_L} \\
&\quad + \left(8\widetilde{p_2}^2+4\widetilde{p_1}\widetilde{p_2}-8\widetilde{p_2}\right)\sqrt{N_C} + (12\widetilde{p_2}+12\widetilde{p_1})\frac{N_C}{\sqrt{N_C}} + (12\widetilde{p_1}+12\widetilde{p_2})\frac{N_L}{\sqrt{N_L}}\Bigg] \\
&\quad \times c_{P,Q}{N_L}^{(P-\widetilde{p_1})/2}{N_R}^{(Q-q)/2}{N_C}^{(1-\widetilde{p_2})/2} \\
&\leq \Bigg[(4\widetilde{p_1}+4\widetilde{p_2}) + \left(8(P+1)^2+4P(P+1)-8\widetilde{p_1}\right)\sqrt{N_L} \\
&\quad + \left(8P^2+4P(P+1)-8\widetilde{p_2}\right)\sqrt{N_C} + (12(P+1)+12P)\frac{N_C}{\sqrt{N_C}} \\
&\quad + (12(P+1)+12P)\frac{N_L}{\sqrt{N_L}}\Bigg] \\
&\quad \times c_{P,Q}{N_L}^{(P-\widetilde{p_1})/2}{N_R}^{(Q-q)/2}{N_C}^{(1-\widetilde{p_2})/2} \\
&\leq (24P^2+72P+32)\times c_{P,Q}{N_L}^{(P+1-\widetilde{p_1})/2}{N_R}^{(Q-q)/2}{N_C}^{(1-\widetilde{p_2})/2} \\
&\leq c_{P+1,Q}{N_L}^{(P+1-\widetilde{p_1})/2}{N_R}^{(Q-q)/2}{N_C}^{(1-\widetilde{p_2})/2},
\end{aligned}
\tag{B.26}
$$

where, for example, the term $\propto 4\widetilde{p_1}\widetilde{p_2}\sqrt{N_L}$ comes from the contribution of Eq. (B.17), with the extra factor $\sqrt{N_L}$ due to $\widetilde{p_1} = p_1 + 1$. The induction hypothesis, Eq. (B.5), is thus satisfied for $P+1, Q$ by choosing $c_{P+1,Q} \geq (24P^2+72P+32)c_{P,Q}$.

Now, we use this to show a bound on the commutator norm. Since different Pauli strings

are orthogonal, the Frobenius norm is bounded as

$$
\begin{aligned}
&\|[H_{\eta_{P+Q}}, \cdots, [H_{\eta_2}, H_{\eta_1}]]\|_{\mathrm{F}}^2 \\
&\leq \sum_{p_1=0}^{P} \sum_{p_2=1}^{P+1-p_1} \sum_{q=0}^{Q} \sum_{l_1<\cdots<l_{p_1}} \sum_{c_1<\cdots<c_{p_2}} \sum_{r_1<r_2<\cdots<r_q} \|A^{\alpha_1\cdots\alpha_{p_1},\beta_1\cdots\beta_q,\gamma_1\cdots\gamma_{p_2}}_{l_1\cdots l_{p_1},c_1\cdots c_{p_2},r_1\cdots r_q;P+1,Q}\|^2 && \text{(B.27)} \\
&\leq \sum_{p_1=0}^{P} \sum_{p_2=1}^{P+1-p_1} \sum_{q=0}^{Q} \sum_{l_1<\cdots<l_{p_1}} \sum_{c_1<\cdots<c_{p_2}} \sum_{r_1<r_2<\cdots<r_q} c_{P+1,Q}^2 {N_L}^{P-p_1} {N_R}^{Q-q} {N_C}^{1-p_2} && \text{(B.28)} \\
&\leq \sum_{p_1=0}^{P} \sum_{p_2=1}^{P+1-p_1} \sum_{q=0}^{Q} \binom{N_L}{p_1}\binom{N_C}{p_2}\binom{N_R}{q} c_{P+1,Q}^2 {N_L}^{P-p_1} {N_R}^{Q-q} {N_C}^{1-p_2} && \text{(B.29)} \\
&\leq \sum_{p_1=0}^{P} \sum_{p_2=1}^{P+1-p_1} \sum_{q=0}^{Q} \frac{N_L^{p_1}}{p_1!}\frac{N_C^{p_2}}{p_2!}\frac{N_R^{q}}{q!} c_{P+1,Q}^2 {N_L}^{P-p_1} {N_R}^{Q-q} {N_C}^{1-p_2} && \text{(B.30)} \\
&\leq (f(P,Q))^2 \, {N_L}^{P} {N_R}^{Q} N_C, && \text{(B.31)}
\end{aligned}
$$

where $f(P,Q)$ does not depend on N_L, N_C, N_R. □

Now we are ready to prove Lemma 3.12, which we restate here.

Lemma B.1.2. *Consider a graph with a vertex bottleneck. Let H be any time-independent Hamiltonian that respects this connectivity. Let $\widetilde{U}$ be the architecture-respecting simulation circuit corresponding to dividing the time-evolution $U = U(H,t)$ into M equal segments, and simulating each by the $(2k)^{\text{th}}$-order Trotter-Suzuki formula. Then, $\widetilde{U}$ has depth $d = 2 \times 5^{k-1} M$ [Zha+22] and there exists a function $g(k)$, only dependent on k, such that*

$$
\|U - \widetilde{U}\|_{\mathrm{F}} \leq g(k) \frac{t^{2k+1}}{M^{2k}} \sqrt{{N_L}^{2k} N_R N_C}. \tag{B.32}
$$

Proof. We use the decomposition in Eq. (3.35) to define $\widetilde{U}$ as the simulation circuit corresponding to dividing the time-evolution by H, $U(H,t)$, into M equal segments, and simulating each by the

$(2k)$th-order Trotter-Suzuki formula. For example, if $k = 1$,

$$\widetilde{U} = \left(e^{-\mathrm{i}(t/2M)H_{LC}} e^{-\mathrm{i}(t/M)H_R} e^{-\mathrm{i}(t/2M)H_{LC}}\right)^M. \tag{B.33}$$

$\widetilde{U}$ consists of $2\times 5^{k-1}M$ gates each acting on either the left and center qubit, or the right and center qubit [Zha+22]. We call $\widetilde{U}$ the circuit approximation. According to Theorem 2 in Ref. [Zha+22] (but using the normalized Frobenius norm),

$$\begin{aligned}
\|U - \widetilde{U}\|_{\mathrm{F}} &\leq 2(2\times 5^{k-1})^{2k+1} M \left(\frac{t}{M}\right)^{2k+1} \sum_{\eta_1,\cdots,\eta_{2k+1}\in\{L,R\}} \|[H_{\eta_{2k+1}},\cdots,[H_{\eta_2},H_{\eta_1}]]\|_{\mathrm{F}}, \\
&\leq 2(4\times 5^{k-1})^{2k+1} M \left(\frac{t}{M}\right)^{2k+1} \max_{\eta_1,\cdots,\eta_{2k+1}\in\{L,R\}} \|[H_{\eta_{2k+1}},\cdots,[H_{\eta_2},H_{\eta_1}]]\|_{\mathrm{F}}.
\end{aligned} \tag{B.34}$$

Let $f^*_{2k+1} \coloneqq \max_{P,Q} f(P,Q)$ s.t. $P+Q = 2k+1$. Then Lemma B.1.1 implies the distance is bounded by

$$\|U - \widetilde{U}\|_{\mathrm{F}} \leq 2(4\times 5^{k-1})^{2k+1} M \left(\frac{t}{M}\right)^{2k+1} \sqrt{f^*_{2k+1}} \sqrt{{N_L}^{2k} N_R N_C}, \tag{B.35}$$

which completes the proof with $g(k) = 2(4\times 5^{k-1})^{2k+1}\sqrt{f^*_{2k+1}}$. □

Appendix C: Appendices for Chapter 4

C.1 Finite Time-Averaged Expectation Values

In this appendix, we prove Lemma 4.44. First, let us recall the lemma:

Lemma 4.44. *Consider the time-averaged expectation value:*

$$\bar{\mathcal{A}}_N(\tau) = \frac{1}{\tau}\int_0^\tau dt \langle\psi_0|e^{iHt}\mathcal{A}_N e^{-iHt}|\psi_0\rangle. \tag{4.93}$$

Then, the following bound between the infinite time and finite time expectation values holds:

$$|\bar{\mathcal{A}}_N(\tau) - \bar{\mathcal{A}}_N| = O\left(\frac{T^2}{\tau}\right). \tag{4.94}$$

Proof. In the finite time case, we can no longer use the integration property of $e^{-i(\lambda_i-\lambda_j)t}$. Instead we write:

$$\begin{aligned}\bar{\mathcal{A}}_N(\tau) &= \frac{1}{\tau}\int_0^\tau dt \sum_{ij} e^{-i(\lambda_i-\lambda_j)} c_i^* c_j \langle\lambda_i|\mathcal{A}_N|\lambda_j\rangle \\ &= \sum_{ij}\delta(\lambda_i-\lambda_j)c_i^*c_j\langle\lambda_i|A|\lambda_j\rangle + \frac{1}{\tau}\sum_{ij}\int_{[\tau_{ij}]} dt e^{-i(\lambda_i-\lambda_j)}c_i^*c_j\langle\lambda_i|\mathcal{A}_N|\lambda_j\rangle,\end{aligned} \tag{C.1}$$

where for a particular i, j we define $[\tau_{ij}]$ as the largest multiple of the period of $e^{-i(\lambda_i-\lambda_j)}$ which

is less than τ. We have used that $e^{-i(\lambda_i-\lambda_j)}$ integrates to zero if the integration is done over a multiple of the oscillatory period. We now want to bound the second term. Using the fact there are at most T eigenstates, we get:

$$\begin{aligned}\frac{1}{\tau}\sum_{ij}\int_{[\tau_{ij}]}^{\tau} dt e^{-i(\lambda_i-\lambda_j)}c_i^* c_j\langle\lambda_i|\mathcal{A}_N|\lambda_j\rangle &\leq \frac{\|A\|}{\tau}\int_{[\tau_{ij}]}^{\tau} dt e^{-i(\lambda_i-\lambda_j)}c_i^* c_j\langle\lambda_i||\lambda_j\rangle \\ &= \frac{\|A\|}{\tau}\int_{[\tau_{ij}]}^{\tau} dt\langle\psi(t)||\psi(t)\rangle \\ &\leq \frac{\|A\|}{\tau}(\tau-[\tau_{ij}]) \end{aligned} \tag{C.2}$$

We now note that $(\tau-[\tau_{ij}])$ must be less than a single period of the exponential $e^{-i(\lambda_i-\lambda_j)t}$, and hence $(\tau-[\tau_{ij}]) \leq \max_{ij} 1/(\lambda_j-\lambda_i)$, where $\max_{ij} 1/(\lambda_j-\lambda_i)$ is the maximum period of any of the complex exponentials. From Section 4.7.1, we see that

$$\lambda_k - \lambda_m = \cos\left(\frac{k\pi}{T+1}\right) - \cos\left(\frac{m\pi}{T+1}\right). \tag{C.3}$$

The minimum will be achieved for $m = k+1$, hence:

$$\begin{aligned}|\lambda_k - \lambda_{k+1}| &= \cos\left(\frac{k\pi}{T+1}\right) - \cos\left(\frac{(k+1)\pi}{T+1}\right) \\ &\leq \frac{1}{2}\left(\frac{\pi}{T+1}\right)^2 |[k^2-(k+1)^2]| + O\left(\frac{1}{T^4}\right) \\ &= \frac{1}{2}\left(\frac{\pi}{T+1}\right)^2 (2k+1) + O\left(\frac{1}{T^4}\right) \end{aligned} \tag{C.4}$$

The minimum is achieved for $k = 0$. Hence:

$$\frac{1}{\tau}\sum_{ij}\int_{[\tau_{ij}]}^{\tau} dt e^{-i(\lambda_i-\lambda_j)} c_i^* c_j \langle\lambda_i|\mathcal{A}_N|\lambda_j\rangle \leq \frac{\|A\|}{\tau}(\tau - [\tau_{ij}])$$
$$\leq O\left(\frac{T^2}{\tau}\right). \tag{C.5}$$

□

Bibliography

[AB09] Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009 (cit. on p. 99).

[Aba+19] Dmitry A. Abanin, Ehud Altman, Immanuel Bloch, and Maksym Serbyn. "Colloquium: Many-body localization, thermalization, and entanglement." In: *Reviews of Modern Physics* 91.2 (May 2019), p. 021001 (cit. on pp. 12, 93).

[AC21] Álvaro M. Alhambra and J. Ignacio Cirac. "Locally Accurate Tensor Networks for Thermal States and Time Evolution." In: *PRX Quantum* 2.4 (Nov. 2021), p. 040331 (cit. on pp. 143, 149).

[ACG94] Noga Alon, F. R. K. Chung, and R. L. Graham. "Routing Permutations on Graphs via Matchings." In: *SIAM Journal on Discrete Mathematics* 7.3 (May 1994), pp. 513–530 (cit. on pp. 6, 16, 39, 40, 45, 55, 160).

[AE07] Andris Ambainis and Joseph Emerson. "Quantum t-designs: t-wise Independence in the Quantum World." In: *22nd Annual IEEE Conference on Computational Complexity*. June 2007, pp. 129–140. arXiv: quant-ph/0701126 (cit. on p. 84).

[AG19] Joran van Apeldoorn and András Gilyén. "Improvements in Quantum SDP-Solving with Applications." In: *LIPIcs, Volume 132, ICALP 2019* 132 (2019). arXiv:1804.05058 [quant-ph], 99:1–99:15 (cit. on p. 105).

[ALY23] Chi-Fang (Anthony) Chen, Andrew Lucas, and Chao Yin. "Speed limits and locality in many-body quantum dynamics." en. In: *Reports on Progress in Physics* 86.11 (Sept. 2023), p. 116001 (cit. on pp. 1, 4).

[Ami+08] Luigi Amico, Rosario Fazio, Andreas Osterloh, and Vlatko Vedral. "Entanglement in many-body systems." In: *Reviews of Modern Physics* 80.2 (May 2008), pp. 517–576 (cit. on p. 154).

[Ami+10] J M Amini et al. "Toward scalable ion traps for quantum information processing." en. In: *New Journal of Physics* 12.3 (Mar. 2010), p. 033031 (cit. on p. 2).

[Ape+20] Joran van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf. "Quantum SDP-Solvers: Better upper and lower bounds." In: *Quantum* 4 (Feb. 2020), p. 230 (cit. on p. 134).

[Aru+19] Frank Arute et al. "Quantum supremacy using a programmable superconducting processor." In: *Nature* 574.7779 (Oct. 2019), pp. 505–510 (cit. on pp. 16, 20).

[Aud07] Koenraad M R Audenaert. "A sharp continuity estimate for the von Neumann entropy." In: *Journal of Physics A: Mathematical and Theoretical* 40.28 (June 2007), p. 8127. arXiv: quant-ph/0610146 (cit. on pp. 79, 84).

[Aud14] Koenraad M. R. Audenaert. "Quantum skew divergence." In: *Journal of Mathematical Physics* 55.11 (Nov. 2014), p. 112202. arXiv: 1304.5935 (cit. on p. 62).

[Bak+24] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Ewin Tang. "High-Temperature Gibbs States are Unentangled and Efficiently Preparable." In: *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*. Oct. 2024, pp. 1027–1036 (cit. on pp. 154, 155).

[Bal+24] Shankar Balasubramanian, Sarang Gopalakrishnan, Alexey Khudorozhkov, and Ethan Lake. "Glassy Word Problems: Ultraslow Relaxation, Hilbert Space Jamming, and Computational Complexity." In: *Physical Review X* 14.2 (May 2024), p. 021034 (cit. on p. 150).

[Bap+21] Aniruddha Bapat et al. "Quantum routing with fast reversals." In: *Quantum* 5 (Aug. 2021), p. 533. arXiv: 2103.03264 (cit. on pp. 17, 45).

[Bap+23] Aniruddha Bapat, Andrew M. Childs, Alexey V. Gorshkov, and Eddie Schoute. "Advantages and Limitations of Quantum Routing." In: *PRX Quantum* 4 (1 Feb. 2023), p. 010313 (cit. on pp. 8, 17, 24, 25, 40, 42, 45, 46, 48, 49, 50, 62, 63, 68, 73, 83, 88, 90, 91, 156).

[Bar+23] Sara Bartolucci et al. "Fusion-based quantum computation." en. In: *Nature Communications* 14.1 (Feb. 2023), p. 912 (cit. on p. 2).

[Bäu+24] Elisa Bäumer et al. "Efficient Long-Range Entanglement Using Dynamic Circuits." In: *PRX Quantum* 5.3 (Aug. 2024), p. 030339 (cit. on p. 6).

[Bay+21] Andreas Bayerstadler et al. "Industry quantum computing applications." en. In: *EPJ Quantum Technology* 8.11 (Dec. 2021), p. 25 (cit. on p. 2).

[BCD01] Harry Buhrman, Richard Cleve, and Wim van Dam. "Quantum Entanglement and Communication Complexity." In: *SIAM Journal on Computing* 30.6 (Jan. 2001), pp. 1829–1841. arXiv: quant-ph/9705033 (cit. on pp. 7, 46).

[BCL24] Thiago Bergamaschi, Chi-Fang Chen, and Yunchao Liu. "Quantum Computational Advantage with Constant-Temperature Gibbs Sampling." In: *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*. Oct. 2024, pp. 1063–1085 (cit. on p. 155).

[BCO17] Johannes Bausch, Toby Cubitt, and Maris Ozols. "The Complexity of Translationally Invariant Spin Chains with Low Local Dimension." en. In: *Annales Henri Poincaré* 18.11 (Nov. 2017), pp. 3449–3513 (cit. on p. 111).

[BE21] Nikolas P. Breuckmann and Jens Niklas Eberhardt. "Quantum Low-Density Parity-Check Codes." In: *PRX Quantum* 2 (4 Oct. 2021), p. 040101 (cit. on p. 19).

[Ben+03] C. H. Bennett, A. W. Harrow, D. W. Leung, and J. A. Smolin. "On the capacities of bipartite Hamiltonians and unitary gates." In: *IEEE Transactions on Information Theory* 49.8 (Aug. 2003), pp. 1895–1911. arXiv: quant-ph/0205057 (cit. on pp. 7, 46).

[Ben+96a] Charles H. Bennett, David P. DiVincenzo, John A. Smolin, and William K. Wootters. "Mixed-state entanglement and quantum error correction." In: *Physical Review A* 54.5 (Nov. 1996), pp. 3824–3851. arXiv: quant-ph/9604024 (cit. on pp. 7, 46).

[Ben+96b] Charles H. Bennett et al. "Purification of Noisy Entanglement and Faithful Teleportation via Noisy Channels." In: *Phys. Rev. Lett.* 76 (5 Jan. 1996), pp. 722–725 (cit. on p. 43).

[Ber+14] Dominic W. Berry, Andrew M. Childs, Richard Cleve, Robin Kothari, and Rolando D. Somma. "Exponential improvement in precision for simulating sparse Hamiltonians." In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. May 2014, pp. 283–292. arXiv: 1312.1414 (cit. on p. 88).

[Ber+15] Dominic W. Berry, Andrew M. Childs, Richard Cleve, Robin Kothari, and Rolando D. Somma. "Simulating Hamiltonian Dynamics with a Truncated Taylor Series." In: *Physical Review Letters* 114.9 (Mar. 2015), p. 090502 (cit. on p. 100).

[BGW24] Adam Bouland, Tudor Giurgica-Tiron, and John Wright. "The state hidden subgroup problem and an efficient algorithm for locating unentanglement." In: arXiv:2410.12706 (Oct. 2024). arXiv:2410.12706 [quant-ph] (cit. on p. 154).

[BH20] Niel de Beaudrap and Steven Herbert. "Quantum linear network coding for entanglement distribution in restricted architectures." In: *Quantum* 4 (Nov. 2020), p. 356 (cit. on p. 17).

[BHV06] S. Bravyi, M. B. Hastings, and F. Verstraete. "Lieb-Robinson Bounds and the Generation of Correlations and Topological Quantum Order." In: *Phys. Rev. Lett.* 97 (5 July 2006), p. 050401 (cit. on p. 44).

[BK02] Sergey B. Bravyi and Alexei Yu. Kitaev. "Fermionic Quantum Computation." In: *Annals of Physics* 298.1 (May 2002), pp. 210–226. arXiv: quant-ph/0003137 (cit. on p. 55).

[BKS22] Michael Beverland, Vadym Kliuchnikov, and Eddie Schoute. "Surface code compilation via edge-disjoint paths." In: *PRX Quantum* 3.2 (May 2022), p. 020342. arXiv: 2110.11493 [quant-ph] (cit. on p. 18).

[Blu+22] Dolev Bluvstein et al. "A quantum processor based on coherent transport of entangled atom arrays." en. In: *Nature* 604.7906 (Apr. 2022), pp. 451–456 (cit. on p. 3).

[Blu+24] Dolev Bluvstein et al. "Logical quantum processor based on reconfigurable atom arrays." en. In: *Nature* 626.7997 (Feb. 2024), pp. 58–65 (cit. on p. 3).

[BM11] D. Bruß and C. Macchiavello. "Multipartite entanglement in quantum algorithms." In: *Physical Review A* 83.5 (May 2011), p. 052313. arXiv: 1007.4179 (cit. on pp. 7, 46).

[BO20] Costin Bădescu and Ryan O'Donnell. "Lower Bounds for Testing Complete Positivity and Quantum Separability." en. In: *LATIN 2020: Theoretical Informatics*. Ed. by Yoshiharu Kohayakawa and Flávio Keidi Miyazawa. Cham: Springer International Publishing, 2020, pp. 375–386 (cit. on p. 154).

[Bol+96] J. J. Bollinger, Wayne M. Itano, D. J. Wineland, and D. J. Heinzen. "Optimal frequency measurements with maximally correlated states." In: *Physical Review A* 54.6 (Dec. 1996), R4649–R4652 (cit. on pp. 7, 46).

[Bom+21] Hector Bombin et al. "Interleaving: Modular architectures for fault-tolerant photonic quantum computing." In: arXiv:2103.08612 (Mar. 2021). arXiv:2103.08612 [quant-ph] (cit. on p. 2).

[Bos03] Sougato Bose. "Quantum Communication through an Unmodulated Spin Chain." In: *Physical Review Letters* 91.20 (Nov. 2003), p. 207901 (cit. on pp. 4, 19).

[Bos07] Sougato Bose. "Quantum communication through spin chain dynamics: an introductory overview." In: *Contemporary Physics* 48.1 (Jan. 2007), pp. 13–30 (cit. on p. 4).

[Bou+21] J. Eli Bourassa et al. "Blueprint for a Scalable Photonic Fault-Tolerant Quantum Computer." en-GB. In: *Quantum* 5 (Feb. 2021), p. 392 (cit. on p. 2).

[Bou+24] Adam Bouland et al. "Public-Key Pseudoentanglement and the Hardness of Learning Ground State Entanglement Structure." en. In: *39th Computational Complexity Conference (CCC 2024)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024, 21:1–21:23 (cit. on p. 155).

[Bra+25] Sergey Bravyi, Dongjin Lee, Zhi Li, and Beni Yoshida. "How much entanglement is needed for quantum error correction?" In: *Physical Review Letters* 134.21 (2025), p. 210602. arXiv: 2405.01332 (cit. on pp. 7, 46).

[Bra07] Sergey Bravyi. "Upper bounds on entangling rates of bipartite Hamiltonians." In: *Physical Review A* 76 (5 Nov. 2007), p. 052319. arXiv: 0704.0964 (cit. on pp. 7, 46, 60).

[Bri+09] H. J. Briegel, D. E. Browne, W. Dür, R. Raussendorf, and M. Van den Nest. "Measurement-based quantum computation." en. In: *Nature Physics* 5.1 (Jan. 2009), pp. 19–26 (cit. on p. 6).

[Bri+98] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller. "Quantum Repeaters: The Role of Imperfect Local Operations in Quantum Communication." In: *Phys. Rev. Lett.* 81 (26 Dec. 1998), pp. 5932–5935 (cit. on p. 17).

[Bri17] Stephen Brierley. "Efficient Implementation of Quantum Circuits with Limited Qubit Interactions." In: *QIC* 17.13&14 (Nov. 2017), pp. 1096–1104. arXiv: 1507.04263 [quant-ph] (cit. on p. 41).

[Bri95] J. Bricmont. "Science of Chaos or Chaos in Science?" In: *Annals of the New York Academy of Sciences* 775.1 (1995), pp. 131–175. arXiv: chao-dyn/9603009 (cit. on p. 92).

[Bru+19a] C. D. Bruzewicz, R. McConnell, J. Stuart, J. M. Sage, and J. Chiaverini. "Dual-species, multi-qubit logic primitives for Ca+/Sr+ trapped-ion crystals." In: *NPJ Quantum Inf.* 5.1 (Nov. 2019), pp. 1–10 (cit. on pp. 18, 20).

[Bru+19b] Colin D. Bruzewicz, John Chiaverini, Robert McConnell, and Jeremy M. Sage. "Trapped-ion quantum computing: Progress and challenges." In: *Applied Physics Reviews* 6.2 (May 2019), p. 021314. arXiv: 1904.04178 (cit. on p. 45).

[BSW25] Elisa Bäumer, David Sutter, and Stefan Woerner. "Approximate Quantum Fourier Transform in Logarithmic Depth on a Line." In: arXiv:2504.20832 (Apr. 2025). arXiv:2504.20832 [quant-ph] (cit. on p. 6).

[BT14] Nikolas P Breuckmann and Barbara M Terhal. "Space-time circuit-to-Hamiltonian construction and its applications." en. In: *Journal of Physics A: Mathematical and Theoretical* 47.19 (Apr. 2014), p. 195304 (cit. on p. 111).

[Bun79] L. A. Bunimovich. "On the ergodic properties of nowhere dispersing billiards." en. In: *Communications in Mathematical Physics* 65.3 (Oct. 1979), pp. 295–312 (cit. on p. 92).

[BV97] Ethan Bernstein and Umesh Vazirani. "Quantum Complexity Theory." In: *SIAM Journal on Computing* 26.5 (Oct. 1997), pp. 1411–1473 (cit. on p. 145).

[BW25] Elisa Bäumer and Stefan Woerner. "Measurement-based long-range entangling gates in constant depth." In: *Physical Review Research* 7.2 (May 2025), p. 023120 (cit. on p. 6).

[BZZ24] Adam Bouland, Chenyi Zhang, and Zixin Zhou. "On the hardness of learning ground state entanglement of geometrically local Hamiltonians." In: arXiv:2411.04353 (Nov. 2024). arXiv:2411.04353 [quant-ph] (cit. on pp. 154, 155).

[Cam+25] Steve Campbell et al. "Roadmap on Quantum Thermodynamics." In: arXiv:2504.20145 (Apr. 2025). arXiv:2504.20145 [quant-ph] (cit. on p. 93).

[CGJ19] Shantanav Chakraborty, András Gilyén, and Stacey Jeffery. "The Power of Block-Encoded Matrix Powers: Improved Regression Techniques via Faster Hamiltonian Simulation." In: *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*. Vol. 132. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019, 33:1–33:14 (cit. on p. 100).

[Chi+21] Andrew M. Childs, Yuan Su, Minh C. Tran, Nathan Wiebe, and Shuchen Zhu. "Theory of Trotter Error with Commutator Scaling." In: *Physical Review X* 11 (1 Feb. 2021), p. 011020. arXiv: 1912.08854 (cit. on pp. 50, 119).

[Chr+05] Matthias Christandl et al. "Perfect transfer of arbitrary states in quantum spin networks." In: *Phys. Rev. A* 71 (3 Mar. 2005), p. 032312 (cit. on p. 19).

[Chu96] Fan Chung. *Spectral Graph Theory*. American Mathematical Society, Dec. 1996 (cit. on pp. 16, 39).

[CKG23] Chi-Fang Chen, Michael J. Kastoryano, and András Gilyén. "An efficient and exact noncommutative quantum Gibbs sampler." In: arXiv:2311.09207 (Nov. 2023). arXiv:2311.09207 [quant-ph] (cit. on p. 155).

[CKS17] Andrew M. Childs, Robin Kothari, and Rolando D. Somma. "Quantum algorithm for systems of linear equations with exponentially improved dependence on precision." In: *SIAM Journal on Computing* 46.6 (Jan. 2017). arXiv:1511.02306 [quant-ph], pp. 1920–1950 (cit. on pp. 100, 128).

[Cle+98] R. Cleve, A. Ekert, C. Macchiavello, and M. Mosca. "Quantum algorithms revisited." In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 454.1969 (Jan. 1998), pp. 339–354 (cit. on p. 102).

[CLY23] Chi-Fang (Anthony) Chen, Andrew Lucas, and Chao Yin. "Speed limits and locality in many-body quantum dynamics." In: *Reports on Progress in Physics* 86.11 (Sept. 2023), p. 116001. arXiv: 2303.07386 (cit. on pp. 49, 50, 87).

[CM06] Nikolai Chernov and Roberto Markarian. *Chaotic Billiards*. en. Vol. 127. Mathematical Surveys and Monographs. American Mathematical Society, July 2006 (cit. on p. 92).

[Cow+19] Alexander Cowtan et al. "On the Qubit Routing Problem." In: *14th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2019)*. Vol. 135. Leibniz International Proceedings in Informatics (LIPIcs). 2019, 5:1–5:32. arXiv: 1902.08091 (cit. on pp. 16, 41, 45).

[CPW15] Toby S. Cubitt, David Perez-Garcia, and Michael M. Wolf. "Undecidability of the spectral gap." en. In: *Nature* 528.7581 (Dec. 2015), pp. 207–211 (cit. on p. 93).

[CS16] Anirban Narayan Chowdhury and Rolando D. Somma. "Quantum algorithms for Gibbs sampling and hitting-time estimation." In: arXiv:1603.02940 (Mar. 2016). arXiv:1603.02940 [quant-ph] (cit. on p. 134).

[CSU19] Andrew M. Childs, Eddie Schoute, and Cem M. Unsal. "Circuit Transformations for Quantum Architectures." In: *14th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2019)*. Vol. 135. Leibniz International Proceedings in Informatics (LIPIcs). 2019, 3:1–3:24. arXiv: 1902.09102 (cit. on pp. 16, 45).

[Cub+03] T. S. Cubitt, F. Verstraete, W. Dür, and J. I. Cirac. "Separable States Can Be Used To Distribute Entanglement." In: *Physical Review Letters* 91 (3 July 2003), p. 037902. arXiv: quant-ph/0302168 (cit. on pp. 7, 46, 49, 90).

[CVC05] T. S. Cubitt, F. Verstraete, and J. I. Cirac. "Entanglement flow in multipartite systems." In: *Physical Review A* 71.5 (May 2005), p. 052308. arXiv: quant-ph/0404179 (cit. on pp. 7, 46, 49, 51, 82).

[DAl+16] Luca D'Alessio, Yariv Kafri, Anatoli Polkovnikov, and Marcos Rigol. "From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics." In: *Advances in Physics* 65.3 (May 2016), pp. 239–362 (cit. on p. 150).

[DBT21] Nicolas Delfosse, Michael E. Beverland, and Maxime A. Tremblay. *Bounds on stabilizer measurement circuits and obstructions to local implementations of quantum LDPC codes*. Sept. 2021. arXiv: 2109.14599 [quant-ph] (cit. on p. 19).

[Den+02] Eric Dennis, Alexei Kitaev, Andrew Landahl, and John Preskill. "Topological quantum memory." In: *Journal of Mathematical Physics* 43.9 (Sept. 2002), pp. 4452–4505 (cit. on p. 43).

[Deu18] Joshua M. Deutsch. "Eigenstate thermalization hypothesis." en. In: *Reports on Progress in Physics* 81.8 (July 2018), p. 082001 (cit. on pp. 11, 12, 92, 93).

[Deu91] J. M. Deutsch. “Quantum statistical mechanics in a closed system.” In: *Physical Review A* 43.4 (Feb. 1991), pp. 2046–2049 (cit. on pp. 11, 93).

[Dev+24] Dhruv Devulapalli, Eddie Schoute, Aniruddha Bapat, Andrew M. Childs, and Alexey V. Gorshkov. “Quantum routing with teleportation.” In: *Physical Review Research* 6.3 (Sept. 2024), p. 033313. arXiv: 2204.04185 (cit. on p. 45).

[DK20] C. M. Da Fonseca and V. Kowalenko. “Eigenpairs of a family of tridiagonal matrices: three decades later.” en. In: *Acta Mathematica Hungarica* 160.2 (Apr. 2020), pp. 376–389 (cit. on p. 112).

[DRC17] C. L. Degen, F. Reinhard, and P. Cappellaro. “Quantum sensing.” In: *Reviews of Modern Physics* 89.3 (July 2017), p. 035002. arXiv: 1611.02427 (cit. on pp. 7, 46).

[Dür+01] W. Dür, G. Vidal, J. I. Cirac, N. Linden, and S. Popescu. “Entanglement Capabilities of Nonlocal Hamiltonians.” In: *Physical Review Letters* 87.13 (Sept. 2001), p. 137901. arXiv: quant-ph/0006034 [quant-ph] (cit. on pp. 7, 46).

[Dut+07] M. V. Gurudev Dutt et al. “Quantum Register Based on Individual Electronic and Nuclear Spin Qubits in Diamond.” In: *Science* 316.5829 (June 2007), pp. 1312–1316 (cit. on pp. 18, 20).

[Ehr+22] Adam Ehrenberg, Abhinav Deshpande, Christopher L. Baldwin, Dmitry A. Abanin, and Alexey V. Gorshkov. “Simulation Complexity of Many-Body Localized Systems.” In: arXiv:2205.12967 (May 2022). arXiv:2205.12967 [quant-ph] (cit. on p. 150).

[EJ98] Artur Ekert and Richard Jozsa. “Quantum algorithms: entanglement–enhanced information processing.” In: *Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences* 356.1743 (Aug. 1998), pp. 1769–1782. arXiv: 9803072 (cit. on pp. 7, 46).

[Eld+20] Zachary Eldredge et al. “Entanglement bounds on the performance of quantum computing architectures.” In: *Phys. Rev. Research* 2 (3 Aug. 2020), p. 033316 (cit. on p. 18).

[Fey82] Richard P. Feynman. “Simulating physics with computers.” en. In: *International Journal of Theoretical Physics* 21.6 (June 1982), pp. 467–488 (cit. on p. 2).

[Fey86] Richard P. Feynman. “Quantum mechanical computers.” en. In: *Foundations of Physics* 16.6 (June 1986), pp. 507–531 (cit. on p. 110).

[Fri+22] Aaron J Friedman, Chao Yin, Yifan Hong, and Andrew Lucas. “Locality and error correction in quantum dynamics with measurement.” In: *arXiv preprint arXiv:2206.09929* (2022) (cit. on p. 19).

[GAE07] D. Gross, K. Audenaert, and J. Eisert. “Evenly distributed unitaries: On the structure of unitary designs.” In: *Journal of Mathematical Physics* 48.5 (May 2007), p. 052104. arXiv: quant-ph/0611002 (cit. on p. 84).

[GC99] Daniel Gottesman and Isaac L. Chuang. “Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations.” In: *Nature* 402.6760 (Nov. 1999), pp. 390–393 (cit. on p. 22).

[GE16] Christian Gogolin and Jens Eisert. “Equilibration, thermalisation, and the emergence of statistical mechanics in closed quantum systems.” en. In: *Reports on Progress in Physics* 79.5 (Apr. 2016), p. 056001 (cit. on pp. 11, 12, 93).

[Gha10] Sevag Gharibian. “Strong NP-hardness of the quantum separability problem.” In: *Quantum Info. Comput.* 10.3 (Mar. 2010), pp. 343–360 (cit. on p. 154).

[Gil+19] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. “Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics.” In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. June 2019, pp. 193–204. arXiv: 1806.01838 [quant-ph] (cit. on pp. 100, 101, 102, 134).

[GLM04] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. “Quantum-Enhanced Measurements: Beating the Standard Quantum Limit.” In: *Science* 306.5700 (Nov. 2004), pp. 1330–1336. arXiv: quant-ph/0412078 (cit. on pp. 7, 46).

[GME11] Christian Gogolin, Markus P. Müller, and Jens Eisert. “Absence of Thermalization in Nonintegrable Systems.” In: *Physical Review Letters* 106.4 (Jan. 2011), p. 040401 (cit. on p. 11).

[GT07] Nicolas Gisin and Rob Thew. “Quantum communication.” In: *Nature Photonics* 1.3 (Mar. 2007), pp. 165–171. arXiv: quant-ph/0703255 [quant-ph] (cit. on pp. 7, 46).

[Gut+15] Gus Gutoski, Patrick Hayden, Kevin Milner, and Mark M. Wilde. “Quantum Interactive Proofs and the Complexity of Separability Testing.” In: *Theory of Computing* 11.3 (Mar. 2015), pp. 59–103 (cit. on p. 154).

[Har66] L.H. Harper. “Optimal numberings and isoperimetric problems on graphs.” In: *Journal of Combinatorial Theory* 1.3 (1966), pp. 385–393 (cit. on p. 40).

[Her20] Steven Herbert. “On the depth overhead incurred when running quantum algorithms on near-term quantum computers with limited qubit connectivity.” In: *QIC* 20.9&10 (Aug. 2020), pp. 787–806. arXiv: 1805.12570v5 [quant-ph] (cit. on pp. 18, 19, 43).

[HH23] Aram W. Harrow and Yichen Huang. “Thermalization without eigenstate thermalization.” In: arXiv:2209.09826 (Nov. 2023). arXiv:2209.09826 [cond-mat] (cit. on p. 11).

[Hon+24] Yifan Hong, Matteo Marinelli, Adam M. Kaufman, and Andrew Lucas. “Long-range-enhanced surface codes.” In: *Phys. Rev. A* 110 (2 Aug. 2024), p. 022607 (cit. on p. 19).

[Hor+09] Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki. “Quantum entanglement.” In: *Reviews of Modern Physics* 81.2 (June 2009), pp. 865–942. arXiv: quant-ph/0702225 (cit. on pp. 7, 46).

[HPE19] F. Hahn, A. Pappa, and J. Eisert. “Quantum network routing and local complementation.” In: *npj Quantum Information* 5.1 (Sept. 2019), pp. 1–7. arXiv: 1805.04559 (cit. on p. 45).

[Hua15] Yichen Huang. "Efficient simulation of many-body localized systems." In: arXiv:1508.04756 (Aug. 2015). arXiv:1508.04756 [cond-mat] (cit. on p. 150).

[JL03] Richard Jozsa and Noah Linden. "On the role of entanglement in quantum-computational speed-up." In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 459.2036 (Aug. 2003), pp. 2011–2032. arXiv: quant-ph/0201143 (cit. on pp. 7, 46).

[Joz98] Richard Jozsa. "Entanglement and Quantum Computation." In: *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose*. Ed. by S A Huggett, L J Mason, K P Tod, S T Tsou, and N M J Woodhouse. Oxford University Press, Apr. 1998, pp. 369–379. arXiv: 9707034 (cit. on pp. 7, 46).

[JW28] P. Jordan and E. Wigner. "Über das Paulische Äquivalenzverbot." In: *Zeitschrift für Physik* 47.9 (Sept. 1928), pp. 631–651 (cit. on p. 66).

[Kau+20] V. Kaushal et al. "Shuttling-based trapped-ion quantum information processing." In: *AVS Quantum Science* 2.1 (Mar. 2020), p. 014101 (cit. on p. 2).

[KIH14] Hyungwon Kim, Tatsuhiko N. Ikeda, and David A. Huse. "Testing whether all eigenstates obey the eigenstate thermalization hypothesis." In: *Physical Review E* 90.5 (Nov. 2014), p. 052105 (cit. on p. 93).

[Kja+20] Morten Kjaergaard et al. "Superconducting Qubits: Current State of Play." In: *Annual Review of Condensed Matter Physics* 11 (2020), pp. 369–395. arXiv: 1905.13641 (cit. on pp. 2, 45).

[KL97] Emanuel Knill and Raymond Laflamme. "Theory of quantum error-correcting codes." In: *Physical Review A* 55.2 (Feb. 1997), pp. 900–911. arXiv: 9604034 (cit. on pp. 7, 46).

[Kow19] Emmanuel Kowalski. *An introduction to expander graphs*. Paris: Société Mathématique de France, 2019 (cit. on p. 160).

[KSV02] A. Kitaev, A. Shen, and M. Vyalyi. *Classical and Quantum Computation*. en. Vol. 47. Graduate Studies in Mathematics. American Mathematical Society, May 2002 (cit. on pp. 102, 111).

[KWW06] Toshiya Kinoshita, Trevor Wenger, and David S. Weiss. "A quantum Newton's cradle." en. In: *Nature* 440.70867086 (Apr. 2006), pp. 900–903 (cit. on pp. 12, 93).

[Lan+15] Tim Langen et al. "Experimental observation of a generalized Gibbs ensemble." In: *Science* 348.6231 (Apr. 2015), pp. 207–211 (cit. on p. 93).

[LC19] Guang Hao Low and Isaac L. Chuang. "Hamiltonian Simulation by Qubitization." In: *Quantum* 3 (July 2019), p. 163 (cit. on p. 100).

[LD98] Daniel Loss and David P. DiVincenzo. "Quantum computation with quantum dots." In: *Phys. Rev. A* 57 (1 Jan. 1998), pp. 120–126 (cit. on pp. 18, 20).

[LDX19] Gushu Li, Yufei Ding, and Yuan Xie. "Tackling the Qubit Mapping Problem for NISQ-Era Quantum Devices." In: *Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems*. Apr. 2019, pp. 1001–1014. arXiv: 1809.02573 (cit. on p. 45).

[Lee+22] Yuan Lee, Eric Bersin, Axel Dahlberg, Stephanie Wehner, and Dirk Englund. "A quantum router architecture for high-fidelity entanglement flows in quantum networks." In: *npj Quantum Information* 8.1 (June 2022), p. 75 (cit. on p. 43).

[LLY10] Wei-Tian Li, Linyuan Lu, and Yiting Yang. "Routing Numbers of Cycles, Complete Bipartite Graphs, and Hypercubes." In: *SIAM J. Discrete Math.* 24.4 (Jan. 2010), pp. 1482–1494 (cit. on p. 40).

[LN22] Daniel Litinski and Naomi Nickerson. "Active volume: An architecture for efficient fault-tolerant quantum computers with limited non-local connections." In: arXiv:2211.15465 (Nov. 2022). arXiv:2211.15465 [quant-ph] (cit. on p. 3).

[LR72] Elliott H. Lieb and Derek W. Robinson. "The finite group velocity of quantum spin systems." In: *Commun. Math. Phys* 28.3 (Sept. 1972), pp. 251–257 (cit. on pp. 3, 4, 17).

[LS25] Eli M. Levenson-Falk and Sadman Ahmed Shanto. "A Review of Design Concerns in Superconducting Quantum Circuits." In: arXiv:2411.16967 (Apr. 2025). arXiv:2411.16967 [quant-ph] (cit. on p. 2).

[Mar+16] Michaël Mariën, Koenraad M. R. Audenaert, Karel Van Acoleyen, and Frank Verstraete. "Entanglement Rates and the Stability of the Area Law for the Entanglement Entropy." In: *Communications in Mathematical Physics* 346.1 (Aug. 2016), pp. 35–73. arXiv: 1411.0680 (cit. on pp. 7, 8, 46, 47, 49, 50, 62).

[MK13] C. Monroe and J. Kim. "Scaling the Ion Trap Quantum Processor." In: *Science* 339.6124 (Mar. 2013), pp. 1164–1169 (cit. on pp. 2, 16).

[Moo90] Cristopher Moore. "Unpredictability and undecidability in dynamical systems." In: *Physical Review Letters* 64.20 (May 1990), pp. 2354–2357 (cit. on p. 93).

[Mor17] Kenichi Morita. "Two Small Universal Reversible Turing Machines." en. In: *Advances in Unconventional Computing: Volume 1: Theory*. Ed. by Andrew Adamatzky. Cham: Springer International Publishing, 2017, pp. 221–237 (cit. on p. 109).

[Mos+23] S.A. Moses et al. "A Race-Track Trapped-Ion Quantum Processor." In: *Physical Review X* 13.4 (Dec. 2023), p. 041052 (cit. on p. 2).

[MW05] Chris Marriott and John Watrous. "Quantum Arthur–Merlin games." en. In: *computational complexity* 14.2 (June 2005), pp. 122–152 (cit. on p. 145).

[Nat+21] Paul Nation, Hanhee Paik, Andrew Cross, and Zaira Nazario. *The IBM Quantum heavy hex lattice*. July 2021 (cit. on p. 20).

[NC10] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. en. Dec. 2010 (cit. on p. 1).

[Ott02] Edward Ott. *Chaos in Dynamical Systems*. 2nd ed. Cambridge: Cambridge University Press, 2002 (cit. on p. 92).

[Ozo89] Alfredo M. Ozorio de Almeida. *Hamiltonian Systems: Chaos and Quantization*. Cambridge Monographs on Mathematical Physics. Cambridge: Cambridge University Press, 1989 (cit. on p. 92).

[Pan+19] Mihir Pant et al. “Routing entanglement in the quantum internet.” In: *npj Quantum Information* 5.11 (Mar. 2019), pp. 1–9 (cit. on p. 43).

[Pat+22] Ashlesha Patil, Mihir Pant, Dirk Englund, Don Towsley, and Saikat Guha. “Entanglement generation in a quantum network at distance-independent rate.” In: *npj Quantum Information* 8.11 (May 2022), pp. 1–9 (cit. on p. 43).

[PB21] R.K. Pathria and Paul D. Beale. *Statistical Mechanics*. 4th. Boston: Academic Press, 2021, pp. 30–32 (cit. on pp. 9, 92).

[Pin+21] J. M. Pino et al. “Demonstration of the trapped-ion quantum CCD computer architecture.” en. In: *Nature* 592.7853 (Apr. 2021), pp. 209–213 (cit. on p. 2).

[PK21] Pavel Panteleev and Gleb Kalachev. “Quantum LDPC codes with almost linear minimum distance.” In: *IEEE Transactions on Information Theory* 68.1 (2021), pp. 213–229 (cit. on p. 19).

[Pre18] John Preskill. “Quantum Computing in the NISQ era and beyond.” en-GB. In: *Quantum* 2 (Aug. 2018), p. 79 (cit. on p. 2).

[PS13] Paul Pham and Krysta M. Svore. “A 2D Nearest-Neighbor Quantum Architecture for Factoring in Polylogarithmic Depth.” In: *QIC* 13.11–12 (2013), pp. 937–962 (cit. on pp. 6, 18, 20, 44).

[PSC21] Lorenzo Piroli, Georgios Styliaris, and J. Ignacio Cirac. “Quantum Circuits Assisted by Local Operations and Classical Communication: Transformations and Phases of Matter.” In: *Phys. Rev. Lett.* 127 (22 Nov. 2021), p. 220503 (cit. on pp. 18, 44).

[PSC24] Lorenzo Piroli, Georgios Styliaris, and J. Ignacio Cirac. “Approximating Many-Body Quantum States with Quantum Circuits and Measurements.” In: *Physical Review Letters* 133.23 (Dec. 2024), p. 230401 (cit. on p. 6).

[PSW06] Sandu Popescu, Anthony J. Short, and Andreas Winter. “Entanglement and the foundations of statistical mechanics.” en. In: *Nature Physics* 2.11 (Nov. 2006), pp. 754–758 (cit. on p. 11).

[PT04] Asher Peres and Daniel R. Terno. “Quantum information and relativity theory.” In: *Reviews of Modern Physics* 76.1 (Jan. 2004), pp. 93–123 (cit. on p. 1).

[RDO08] Marcos Rigol, Vanja Dunjko, and Maxim Olshanii. “Thermalization and its mechanism for generic isolated quantum systems.” en. In: *Nature* 452.7189 (Apr. 2008), pp. 854–858 (cit. on p. 93).

[Ren+04] Joseph M. Renes, Robin Blume-Kohout, A. J. Scott, and Carlton M. Caves. “Symmetric informationally complete quantum measurements.” In: *Journal of Mathematical Physics* 45.6 (June 2004), pp. 2171–2180. arXiv: quant-ph/0310075 (cit. on p. 84).

[RFA24] Cambyse Rouzé, Daniel Stilck França, and Álvaro M. Alhambra. “Efficient thermalization and universal quantum computing with quantum Gibbs samplers.” In: arXiv:2403.12691 (Oct. 2024). arXiv:2403.12691 [quant-ph] (cit. on p. 155).

[Rig+07] Marcos Rigol, Vanja Dunjko, Vladimir Yurovsky, and Maxim Olshanii. "Relaxation in a Completely Integrable Many-Body Quantum System: An Ab Initio Study of the Dynamics of the Highly Excited States of 1D Lattice Hard-Core Bosons." In: *Physical Review Letters* 98.5 (Feb. 2007), p. 050405 (cit. on pp. 12, 150).

[Ros13] David J. Rosenbaum. "Optimal Quantum Circuits for Nearest-Neighbor Architectures." In: *8th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2013)*. Vol. 22. Leibniz International Proceedings in Informatics (LIPIcs). 2013, pp. 294–307. arXiv: 1205.0036 (cit. on pp. 18, 43, 45).

[RS12] Marcos Rigol and Mark Srednicki. "Alternatives to Eigenstate Thermalization." In: *Physical Review Letters* 108.11 (Mar. 2012), p. 110601 (cit. on p. 11).

[Rue99] David Ruelle. *Statistical Mechanics: Rigorous Results*. reprint. World Scientific, 1999, pp. 1–5 (cit. on p. 92).

[Shi03] Yu Shi. "Quantum entanglement of identical particles." In: *Physical Review A* 67.2 (Feb. 2003), p. 024301. arXiv: quant-ph/0205069 (cit. on p. 62).

[Sim04] Nándor Simányi. "Proof of the Ergodic Hypothesis for Typical Hard Ball Systems." In: *Annales Henri Poincaré* 5.2 (Apr. 2004), pp. 203–233 (cit. on p. 92).

[Sin70] Yakov G Sinai. "Dynamical systems with elastic reflections." In: *Russian Mathematical Surveys* 25.2 (Apr. 1970), p. 137 (cit. on p. 92).

[SM21] Naoto Shiraishi and Keiji Matsumoto. "Undecidability in quantum thermalization." In: *Nature Communications* 12.1 (2021) (cit. on pp. 12, 93, 94, 108, 109, 113, 148).

[Sre94] Mark Srednicki. "Chaos and quantum thermalization." In: *Physical Review E* 50.2 (Aug. 1994), pp. 888–901 (cit. on pp. 11, 93).

[SV14] Sushant Sachdeva and Nisheeth K. Vishnoi. "Faster Algorithms via Approximation Theory." English. In: *Foundations and Trends® in Theoretical Computer Science* 9.2 (Mar. 2014), pp. 125–210 (cit. on p. 102).

[Swi18] Brian Swingle. "Unscrambling the physics of out-of-time-order correlators." en. In: *Nature Physics* 14.10 (Oct. 2018), pp. 988–990 (cit. on pp. 3, 11).

[Tan+24a] Daniel Bochen Tan, Dolev Bluvstein, Mikhail D. Lukin, and Jason Cong. "Compiling Quantum Circuits for Dynamically Field-Programmable Neutral Atoms Array Processors." en-GB. In: *Quantum* 8 (Mar. 2024), p. 1281 (cit. on p. 3).

[Tan+24b] Nathanan Tantivasadakarn, Ryan Thorngren, Ashvin Vishwanath, and Ruben Verresen. "Long-Range Entanglement from Measuring Symmetry-Protected Topological Phases." In: *Physical Review X* 14.2 (June 2024), p. 021040 (cit. on pp. 6, 18, 44).

[Tas16] Hal Tasaki. "Typicality of Thermal Equilibrium and Thermalization in Isolated Macroscopic Quantum Systems." en. In: *Journal of Statistical Physics* 163.5 (June 2016). Company: Springer Distributor: Springer Institution: Springer Label: Springer number: 5 publisher: Springer US, pp. 937–997 (cit. on pp. 12, 93).

[Ter15] Barbara M. Terhal. "Quantum error correction for quantum memories." In: *Reviews of Modern Physics* 87.2 (Apr. 2015), pp. 307–346 (cit. on p. 6).

[Tur+18] C. J. Turner, A. A. Michailidis, D. A. Abanin, M. Serbyn, and Z. Papić. "Weak ergodicity breaking from quantum many-body scars." en. In: *Nature Physics* 14.77 (July 2018), pp. 745–749 (cit. on pp. 12, 93).

[VC05] F Verstraete and J I Cirac. "Mapping local Hamiltonians of fermions to local Hamiltonians of spins." In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.09 (Sept. 2005), P09012. arXiv: quant-ph/0508353 (cit. on p. 66).

[VKS24] Tan Van Vu, Tomotaka Kuwahara, and Keiji Saito. "Optimal light cone for macroscopic particle transport in long-range systems: A quantum speed limit approach." In: *Quantum* 8 (Sept. 2024), p. 1483. arXiv: 2307.01059 (cit. on p. 91).

[VMV13] Karel Van Acoleyen, Michaël Mariën, and Frank Verstraete. "Entanglement Rates and Area Laws." In: *Physical Review Letters* 111 (17 Oct. 2013), p. 170501. arXiv: 1304.5931 (cit. on pp. 7, 8, 46, 48, 49, 50, 51, 60, 70, 71, 82, 87).

[VR16] Lev Vidmar and Marcos Rigol. "Generalized Gibbs ensemble in integrable lattice models." In: *Journal of Statistical Mechanics: Theory and Experiment* 2016.6 (June 2016), p. 064007 (cit. on p. 93).

[VTV22] Ruben Verresen, Nathanan Tantivasadakarn, and Ashvin Vishwanath. *Efficiently preparing Schrödinger's cat, fractons and non-Abelian topological order in quantum devices*. 2022. arXiv: 2112.03061 [quant-ph] (cit. on pp. 18, 44).

[Wag+23] Friedrich Wagner, Andreas Bärmann, Frauke Liers, and Markus Weissenbäck. "Improving Quantum Computation by Optimized Qubit Routing." In: *Journal of Optimization Theory and Applications* 197.3 (June 2023), pp. 1161–1194. arXiv: 2206.01294 (cit. on p. 45).

[Wat08] John Watrous. "Quantum Computational Complexity." In: arXiv:0804.3401 (Apr. 2008) (cit. on pp. 99, 100).

[Wat19] James D. Watson. "Detailed Analysis of Circuit-to-Hamiltonian Mappings." In: arXiv:1910.01481 (Oct. 2019). arXiv:1910.01481 [quant-ph] (cit. on p. 111).

[Wat99] John Watrous. "Space-Bounded Quantum Complexity." In: *Journal of Computer and System Sciences* 59.2 (Oct. 1999), pp. 281–326 (cit. on pp. 99, 118, 123, 134).

[YC08] Wen-Chyuan Yueh and Sui Sun Cheng. "Explicit eigenvalues and inverses of tridiagonal Toeplitz matrices with four perturbed corners." In: *The ANZIAM Journal* 49.3 (Jan. 2008), pp. 361–387 (cit. on p. 112).

[Yin25] Chao Yin. "Fast and Accurate Greenberger-Horne-Zeilinger Encoding Using All-to-All Interactions." In: *Physical Review Letters* 134.13 (Apr. 2025), p. 130604. arXiv: 2406.10336 (cit. on pp. 89, 91).

[YLS24] Chao Yin, Andrew Lucas, and David T. Stephen. "Frobenius light cone and the shift unitary." In: *Physical Review A* 109 (5 May 2024), p. 052419. arXiv: 2402.07990 (cit. on p. 50).

[YZ25] Pei Yuan and Shengyu Zhang. "Full Characterization of the Depth Overhead for Quantum Circuit Compilation with Arbitrary Qubit Connectivity Constraint." In: *Quantum* 9 (May 2025), p. 1757. arXiv: 2402.02403 (cit. on p. 46).

[Zan02] Paolo Zanardi. “Quantum entanglement in fermionic lattices.” In: *Physical Review A* 65.4 (Mar. 2002), p. 042101. arXiv: quant-ph/0104114 (cit. on p. 62).

[Zha+22] Qi Zhao, You Zhou, Alexander F. Shaw, Tongyang Li, and Andrew M. Childs. “Hamiltonian Simulation with Random Inputs.” In: *Physical Review Letters* 129 (27 Dec. 2022), p. 270502. arXiv: 2111.04773 (cit. on pp. 50, 74, 75, 86, 175, 176).

[Zha99] Louxin Zhang. “Optimal Bounds for Matching Routing on Trees.” In: *SIAM J. Discrete Math.* 12.1 (Jan. 1999), pp. 64–77 (cit. on pp. 16, 25, 26).

[Żuk+93] M. Żukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert. ““Event-ready-detectors” Bell experiment via entanglement swapping.” In: *Physical Review Letters* 71.26 (Dec. 1993). Publisher: American Physical Society, pp. 4287–4290 (cit. on pp. 17, 22).

[ZZC25] Qi Zhao, You Zhou, and Andrew M. Childs. “Entanglement accelerates quantum simulation.” en. In: *Nature Physics* (July 2025), pp. 1–8 (cit. on pp. 7, 46).