

## Analyzing Prospects for Quantum Advantage in Topological Data Analysis

Dominic W. Berry,<sup>1,\*</sup> Yuan Su,<sup>2</sup> Casper Gyurik,<sup>3</sup> Robbie King,<sup>2,4</sup> Joao Basso,<sup>2</sup>  
Alexander Del Toro Barba<sup>2</sup>,<sup>2</sup> Abhishek Rajput<sup>5</sup>,<sup>5</sup> Nathan Wiebe,<sup>5,6</sup> Vedran Dunjko,<sup>3</sup> and  
Ryan Babbush<sup>2,†</sup>

<sup>1</sup>*School of Mathematical and Physical Sciences, Macquarie University, Sydney, NSW 2109, Australia*

<sup>2</sup>*Google Quantum AI, Venice, California 90291, USA*

<sup>3</sup>*applied Quantum algorithms (aQa), Leiden University, Leiden 2300 RA, Netherlands*

<sup>4</sup>*Department of Computing and Mathematical Sciences, Caltech, Pasadena, California 91125, USA*

<sup>5</sup>*Department of Computer Science, University of Toronto, Ontario M5S 2E4, Canada*

<sup>6</sup>*Pacific Northwest National Laboratory, Richland, Washington 99354, USA*



(Received 26 April 2023; revised 26 September 2023; accepted 22 December 2023; published 6 February 2024)

Lloyd *et al.* [Nat. Commun. 7, 10138 (2016)] were first to demonstrate the promise of quantum algorithms for computing Betti numbers, a way to characterize topological features of data sets. Here, we propose, analyze, and optimize an improved quantum algorithm for topological data analysis (TDA) with reduced scaling, including a method for preparing Dicke states based on inequality testing, a more efficient amplitude estimation algorithm using Kaiser windows, and an optimal implementation of eigenvalue projectors based on Chebyshev polynomials. We compile our approach to a fault-tolerant gate set and estimate constant factors in the Toffoli complexity. Our analysis reveals that superquadratic quantum speedups are only possible for this problem when targeting a multiplicative error approximation and the Betti number grows asymptotically. Further, we propose a dequantization of the quantum TDA algorithm that shows that having exponentially large dimension and Betti number are necessary, but insufficient conditions, for superpolynomial advantage. We then introduce and analyze specific problem examples which have parameters in the regime where superpolynomial advantages may be achieved, and argue that quantum circuits with tens of billions of Toffoli gates can solve seemingly classically intractable instances.

DOI: [10.1103/PRXQuantum.5.010319](https://doi.org/10.1103/PRXQuantum.5.010319)

### I. INTRODUCTION

An important outstanding challenge in quantum computing is to find quantum algorithms that provide a significant speedup for practical problems. One area of great interest is quantum machine learning [1]. Early proposals included, for example, principal component analysis [2], and were often based on quantum solution of linear equations [3]. However, it has proven possible to dequantize many of these proposals, indicating that there is at most a polynomial speedup [4,5]. Analysis of the cost taking into account error-correction overhead indicates that more than a quadratic speedup would be needed to provide a useful quantum advantage within quantum error correction [6,7].

An algorithm for topological data analysis proposed by Lloyd *et al.* [8] turned out not to be directly “dequantizable” using the same techniques, raising the question of whether a greater speedup was possible. A simple analysis by Gunn *et al.* [9] contradicted some of the scaling results originally reported by Lloyd *et al.*, and indicated that under certain assumptions there would still only be a quadratic speedup for these algorithms (our analysis agrees with that of Gunn *et al.*). Here we give a far more careful analysis of the complexity, and examine applications which provide better-than-quadratic speedups.

An important goal in data analysis is to extract features of a data set and use them to cluster or classify the data. This data set would be represented as a set of points in some metric space, such as  $\mathbb{R}^n$  with the Euclidean distance function. One approach for the analysis is to convert the point cloud into a graph where the vertices are the given data points and the edges are determined by whether or not pairs of points lie within a chosen distance  $\epsilon$ . This approach can capture features such as connectivity but ignores potential higher-dimensional features, especially if the data points are sampled from

\*Corresponding authors. dominic.berry@mq.edu.au

†ryanbabbush@gmail.com

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

some underlying high-dimensional manifold. Topological data-analysis (TDA) attempts to extract such higher-dimensional global topological features of an underlying data set by applying techniques from the field of algebraic topology, in particular, what is known as simplicial homology.

A simplex is a point, line segment, triangle, or higher-dimensional equivalent, and a simplicial complex is a collection of simplexes. One can form a simplicial complex from the data set with respect to a distance scale  $\epsilon$ , by adding points that are within distance  $2\epsilon$  to simplexes. The Betti number  $\beta_k$  is the number of  $k$ -dimensional holes of the complex. One can determine the Betti number for a chosen range of  $\epsilon$ . Betti numbers, which persist over an appreciable range of the values of  $\epsilon$  are indicative of intrinsic topological features of the data set, as opposed to artifacts that appear at a particular scale and disappear shortly thereafter. The study of such features is referred to as persistent homology.

The classical complexities of algorithms for estimating Betti numbers are typically exponential in  $k$ . That means the computation can be intractable even for a moderate amount of data. That is an important feature for the promise of quantum algorithms, because even fully error-corrected quantum computers with millions of physical qubits are expected to be very limited in data storage. The most promising applications of quantum computers are therefore those involving a limited amount of classical data that needs to be fed into the quantum algorithm as part of the problem specification.

Recent work on quantum TDA algorithms introduced more efficient fermionic representations of the Dirac operator [10] and employed the quantum singular value transformation to implement the kernel projector [11,12]. Some of these techniques have led to significant asymptotic improvements over the original approach, but it is unclear whether they are useful for reducing the fault-tolerant implementation cost for solving problems in practice. Indeed, to the best of our knowledge, no study has been done on the fault-tolerant implementation of quantum TDA algorithms for solving *any* instance of problems of practical interest.

In this work, we give a new algorithm for estimating Betti numbers on a quantum computer. We significantly reduce the cost of fault-tolerant implementation as compared to prior work, as well as estimating the constant factors that are needed to give realistic estimates of gate counts. Specifically, we develop a new method to prepare the initial Dicke states, introduce improved amplitude estimation using Kaiser windows, directly construct the quantum walk operator from block encoding, optimally project onto the kernel of the boundary map, then use the overlap estimation to estimate the kernel dimension of the block-encoded operator, leading to a quadratic improvement in precision over classical sampling. We

also provide the concrete constant factors in the complexity of our algorithm and estimate its fault-tolerant cost, going beyond the asymptotic analyses of all existing work on quantum topological data analysis. Finally, we show that it is possible to construct specific data sets, which have parameters in a regime where quantum TDA would appear to have a significant speedup. In particular, we give examples of a very specific family of problem instances exhibiting the required parameters for the quantum TDA to have a superpolynomial speedup over the naive general classical algorithm, and a more general family of instances that exhibits the required parameters for a quartic speedup. Here, we are comparing to well-studied classical approaches with complexity approximately linear in the possible number of cliques.

We provide a more detailed explanation of the technical background needed to understand Betti numbers in Sec. II. We then provide the improved algorithm and the analysis of its complexity in Sec. III. We use this result to analyze the regimes where large quantum speedups may be expected in Sec. IV. In particular, we consider cases where the Betti number would be large (implying a large quantum speedup) in Sec. IV A, and novel competing classical algorithms in Sec. IV D. We then conclude in Sec. V.

## II. TECHNICAL BACKGROUND

Here we give a more detailed background that is needed to understand the standard approaches for this problem and our contribution. For the technical definitions of the simplicial complex and Betti number, see Appendix A.

### A. Overview of the TDA algorithm and its implementation

In order to analyze the Betti numbers, the points and lines between the points are represented by a graph  $G$ . Then a simplex is represented by a clique in the graph (groups of vertices that are all connected by edges). The  $n$  vertices of the graph  $G$  are represented by  $n$  qubits. That is,  $|1\rangle|0\rangle\cdots|0\rangle$  would represent the first vertex, and  $|0\rangle|0\rangle\cdots|1\rangle|0\rangle$  would represent vertex  $n-1$ . Note that this is a distinct representation from that often used to analyze sparse Hamiltonians, where each computational basis state represents a distinct vertex (so  $n$  qubits would represent  $2^n$  vertices). In the representation here, a computational basis state with more than one  $|1\rangle$  would represent a clique of the graph (a set of vertices with edges connecting every pair of vertices). For example,  $|1\rangle|1\rangle|1\rangle|0\rangle\cdots|0\rangle$  would represent a clique of the first three vertices.

The entire Hilbert space can then be subdivided into subspaces of different Hamming weights. Using  $\mathcal{H}_k$  to denote the space spanned by computational basis states

with Hamming weight  $k$ , we have

$$(\mathbb{C}^2)^{\otimes n} = \bigoplus_{k=0}^n \mathcal{H}_k, \quad (1)$$

where  $\dim(\mathcal{H}_k) = \binom{n}{k}$ . This space includes all states of the various Hamming weights. One can also restrict to only states which represent vertices of cliques of the graph  $G$ . Denoting by  $\mathcal{H}_k^G$  the space spanned by basis states of all  $k$ -cliques of  $G$ , we have  $\mathcal{H}_k^G \subseteq \mathcal{H}_k$ . We also use  $\text{Cl}_k(G)$  to denote the set of bit strings which correspond to  $k$ -cliques of  $G$ .

We define boundary maps  $\partial_k : \mathcal{H}_{k+1} \rightarrow \mathcal{H}_k$  by their actions on the basis states  $|x\rangle \in \mathcal{H}_{k+1}$  as

$$\partial_k |x\rangle := \sum_{i=0}^k (-1)^i |x \setminus (i)\rangle, \quad (2)$$

where  $x \setminus (i)$  means the  $i$ th 1 in the bit string  $x$  is set to 0. We also define  $\partial_k^G : \mathcal{H}_{k+1}^G \rightarrow \mathcal{H}_k^G$  as the restriction of  $\partial_k$  to  $\mathcal{H}_{k+1}^G$ . That is, it gives zero for any  $x$  not representing a  $k+1$ -clique of  $G$ . By definition, we have that both  $\text{im}(\partial_{k+1}^G)$  and  $\text{ker}(\partial_k^G)$  are subspaces of  $\mathcal{H}_k^G$ . But in fact, we have  $\text{im}(\partial_{k+1}^G) \subseteq \text{ker}(\partial_k^G) \subseteq \mathcal{H}_k^G$ , which can be seen from (with  $|x\rangle \in \mathcal{H}_k^G$ )

$$\begin{aligned} \partial_k^G \partial_{k+1}^G |x\rangle &= \sum_{i=0}^{k+1} (-1)^i \partial_k^G |x \setminus (i)\rangle \\ &= \sum_{i=0}^{k+1} (-1)^i \sum_{j=0}^{i-1} (-1)^j |x \setminus (j, i)\rangle \\ &\quad + \sum_{i=0}^{k+1} (-1)^i \sum_{j=i+1}^k (-1)^{j-1} |x \setminus (i, j)\rangle \\ &= \sum_{i=0}^{k+1} (-1)^i \sum_{j=0}^{i-1} (-1)^j |x \setminus (j, i)\rangle \\ &\quad + \sum_{j=0}^{k+1} (-1)^{j-1} \sum_{i=0}^{j-1} (-1)^i |x \setminus (i, j)\rangle \\ &= 0. \end{aligned} \quad (3)$$

Since  $\text{im}(\partial_{k+1}^G)$  is a subspace of  $\text{ker}(\partial_k^G)$ , one can define the quotient space

$$H_k(G) := \text{ker}(\partial_k^G) / \text{im}(\partial_{k+1}^G). \quad (4)$$

This space is called the  $k$ th homology group, and its dimension

$$\beta_k^G := \dim(H_k(G)) = \dim(\text{ker}(\partial_k^G)) - \dim(\text{im}(\partial_{k+1}^G)) \quad (5)$$

is the  $k$ th Betti number. In practice, Betti numbers  $\beta_k^G$  can be used to extract features of the shape of the data modeled by the graph  $G$ , and their estimation is the main problem in the topological data analysis we will consider here. In this work we will be estimating  $\beta_{k-1}^G$  for the  $k-1$ th Betti number, so we can simplify our discussion by considering Hamming weight  $k$ .

To describe our quantum algorithm and its circuit implementation for estimating Betti numbers, we will introduce the Dirac operator  $B_G$ . Specifically, for any graph  $G$  and a fixed value of  $k$ , we define

$$B_G := \begin{bmatrix} 0 & \partial_{k-1}^G & 0 \\ \partial_{k-1}^{G\dagger} & 0 & \partial_k^G \\ 0 & \partial_k^{G\dagger} & 0 \end{bmatrix}, \quad (6)$$

where the blocks indicate the subspaces  $\mathcal{H}_{k-1}^G$ ,  $\mathcal{H}_k^G$ , and  $\mathcal{H}_{k+1}^G$ . Since  $\partial_{k-1}^G \partial_k^G$  gives zero, squaring  $B_G$  yields

$$B_G^2 = \begin{bmatrix} \partial_{k-1}^G \partial_{k-1}^{G\dagger} & 0 & 0 \\ 0 & \partial_{k-1}^{G\dagger} \partial_{k-1}^G + \partial_k^G \partial_k^{G\dagger} & 0 \\ 0 & 0 & \partial_k^{G\dagger} \partial_k^G \end{bmatrix}. \quad (7)$$

It can be seen here that the middle part corresponds to the combinatorial Laplacian [13]

$$\Delta_{k-1}^G = \partial_{k-1}^{G\dagger} \partial_{k-1}^G + \partial_k^G \partial_k^{G\dagger}. \quad (8)$$

It is known that [13–15]

$$\dim(\text{ker}(\Delta_{k-1}^G)) = \beta_{k-1}^G, \quad (9)$$

which provides a convenient way of computing Betti numbers. Since  $B_G$  is Hermitian, the kernel of  $B_G$  and  $B_G^2$  is identical. Therefore, to estimate the Betti number corresponding to a particular graph and a fixed value of  $k$ , it suffices to construct the Dirac operator and compute the dimension of its kernel on the subspace  $\mathcal{H}_k^G$ .

It can be difficult in general to perform topological data analysis on a classical computer due to the high-dimensional nature of the problem, with the dimension increasing exponentially in  $k$ . However, the Dirac operator could be efficiently simulated on a quantum computer, in the sense of solving the Schrödinger equation with the Dirac operator as the Hamiltonian. That indicates exponential speedups are possible, though there are a number of other stages needed for the quantum algorithm. Previous work has provided several approaches for estimating Betti numbers on quantum computers. The stages of these approaches include preparation of a uniformly mixed state, construction of the projector onto the kernel subspace, and estimation of the overlap.

The original approach of Lloyd *et al.* [8] applied amplitude amplification and estimation to prepare the desired initial state in  $\mathcal{H}_k^G$ , starting from a uniform mixture of all Hamming weight  $k$  basis states in  $\mathcal{H}_k$ . Their approach actually produces a superposition over all values of  $k$ , so the success probability of obtaining a specific  $k$  can be quite low; this issue was addressed by later work such as Refs. [9,15]. To construct the projector onto the kernel, they implement Hamiltonian simulation and perform quantum phase estimation on the resulting operator. The Betti number is then estimated as the frequency of zero eigenvalues in the measurement. That is, the algorithm can be summarized as below.

(1) For  $i = 1, \dots, m$ , repeat the following:

(a) Prepare the mixed state

$$\rho_k^G = \frac{1}{\dim(\mathcal{H}_k^G)} \sum_{x \in \text{Cl}_k(G)} |x\rangle\langle x|. \quad (10)$$

(b) Apply quantum phase estimation to the unitary  $e^{iB_G t}$ .

(c) Measure the eigenvalue register to obtain an approximation  $\tilde{\lambda}_i$ .

(2) Output the frequency of zero eigenvalues:

$$\frac{\#\{i, \tilde{\lambda}_i = 0\}}{m}. \quad (11)$$

In this work, we give a new algorithm for estimating Betti numbers on a quantum computer. We provide a number of improvements, which significantly reduce the cost of fault-tolerant implementation. Specifically, we do the following.

- (A) Develop new methods to prepare a mixture of fixed Hamming-weight states with garbage information that have significantly lower fault-tolerant cost. This step is a precursor to preparing the equivalent of Eq. (10).
- (B) Introduce improved amplitude estimation using Kaiser windows to estimate the number of steps of amplitude estimation needed. That is used both for preparation of Eq. (10) and estimation of the frequency of zero eigenvalues.
- (C) Directly construct the quantum walk operator from block encoding, which is more efficient than quantum simulation to give the operator  $e^{iB_G t}$ .
- (D) Project onto the kernel of the boundary map by implementing a Chebyshev polynomial to optimally filter the zero eigenvalues. This is more efficient than phase estimation as in step 1 (b) above.
- (E) Use the overlap estimation to estimate the kernel dimension of the block-encoded operator, leading

to a quadratic improvement in precision over the sampling approach as in step 2 above.

We also provide the concrete constant factors in the complexity of our algorithm and estimate its fault-tolerant cost for solving example problems, going beyond the asymptotic analyses of all prior work on quantum topological data analysis.

Ultimately, the performance of our algorithm (as well as other algorithms from previous work) will depend on several important problem parameters. First, the desired state on which we perform the kernel projector as in Eq. (10) is a uniform mixture of all the  $|\text{Cl}_k(G)|$  basis states in  $\mathcal{H}_k^G$ , where  $|\text{Cl}_k(G)|$  is the number of  $k$ -cliques. In contrast, we start with a uniform mixture of all  $\binom{n}{k}$  basis states in  $\mathcal{H}_k$  [the mixture of Hamming weight  $k$  states referred to in (A) above]. The ratio  $\binom{n}{k}/|\text{Cl}_k(G)|$  will determine the number of amplification steps required in the state preparation. There is potential to improve the efficiency of preparation of the cliques via an improved clique-finding algorithm.

Second, we need to implement a spectral projector that distinguishes the zero eigenvalue from the remaining nonzero eigenvalues of the Dirac operator. That is the projection referred to in (D) above, which replaces the phase estimation from prior work. The cost of implementing such a projector will depend on the spectral gap of the Dirac operator. Third, the output of the quantum TDA algorithm [given by the overlap estimation in (E) above] will not be the actual Betti number  $\beta_{k-1}^G$  but instead a normalized version  $\beta_{k-1}^G/|\text{Cl}_k(G)|$ . In order to estimate the Betti number to some additive precision, we need to increase the complexity by a factor that depends on  $|\text{Cl}_k(G)|$ , with the result that the complexity would roughly scale as  $\sqrt{\binom{n}{k}}$ .

An alternative scenario is that a fixed relative error is required; that is, the ratio of the uncertainty in the Betti number to the Betti number. Then the complexity would roughly scale as  $\sqrt{\binom{n}{k}/\beta_{k-1}^G}$ , as we show in Sec. III. This means that significant speedups can be provided in cases where the Betti number  $\beta_{k-1}^G$  is large, and we provide examples of such graphs in Sec. IV.

Our overall complexity may be summarized as in the following lemma.

*Lemma 1 (Total complexity).*—The complexity of estimating to relative error  $r$  the Betti number  $\beta_{k-1}^G$  of graph  $G$  with  $n$  vertices may be approximated as, for two different methods

$$\frac{\ln(1/\delta_2)}{r_2} \sqrt{\frac{|\text{Cl}_k(G)|}{\beta_{k-1}^G}} \left[ \frac{\pi}{2} \sqrt{\frac{\binom{n}{k}}{|\text{Cl}_k(G)|}} (6|E| + n \log^2 n) + \frac{n}{\lambda_{\min}} \ln \left( \frac{4|\text{Cl}_k(G)|}{r_3 \beta_{k-1}^G} \right) (6|E| + 5n) \right], \quad (12)$$

$$\frac{\ln(1/\delta_2)}{r_2} \sqrt{\frac{|\text{Cl}_k(G)|}{\beta_{k-1}^G}} \left[ \frac{\pi}{2} \sqrt{\frac{n^k/k!}{|\text{Cl}_k(G)|}} (6|E| + 2kn) + \frac{n}{\lambda_{\min}} \ln \left( \frac{4|\text{Cl}_k(G)|}{r_3 \beta_{k-1}^G} \right) (6|E| + 5n) \right], \quad (13)$$

with probability of failure  $\delta = \delta_1 + \delta_2$ , where  $r = r_1 + r_2 + r_3$ ,  $|\text{Cl}_k(G)|$  is the number of  $k$ -cliques, and it is assumed we are given a classical database of edges of the graph. In the case where we are instead given a database of missing edges, then  $6|E|$  is replaced with  $4|E^C|$  in the above expressions.

See Sec. III F for the explanation of this total complexity.

### B. Complexity classes of TDA

Linear-algebraic applications of quantum computing have led to numerous suggestions of how various types of machine-learning subroutines could be implemented on a quantum computer with superpolynomial speedups over their classical counterparts. Many of these methods were in the end shown to suffice only for at most polynomial speedups, due to the randomized “dequantizations” of Tang and others [4,5,16]. The algorithm of Lloyd *et al.* [8], however, turned out not to be directly “dequantizable” using similar techniques, raising the question of whether more robust complexity-theoretic quantum-classical separations can be proven. The current landscape on this topic is somewhat involved.

In general, we have a number of discrepancies between the computational problems in ordinary TDA applications and the computational problems for which we have certain complexity-theoretic insights. In ordinary TDA applications one is typically concerned with the computation of the exact count of zero eigenvalues of combinatorial Laplacians. By the result of Ref. [17]—which shows that deciding if a combinatorial Laplacian has a trivial or non-trivial kernel (i.e., Betti number zero or nonzero) is  $\text{QMA}_1$ -hard—this problem is likely beyond what is efficient even for quantum computers in the worst case. This observation goes in line with classical bodies of work showing that exact computations of Betti numbers are NP-hard [18], and that it can even be PSPACE-hard for more involved topological spaces (i.e., so-called *algebraic varieties*) [19].

From the perspective of the types of problems quantum algorithms may be efficient for, one could attempt a few relaxations of the problem. First, it may be fruitful to relax the TDA problem with respect to the quantity estimated. Specifically, instead of the number of exactly zero eigenvalues, one could relax it and count the number of “small” eigenvalues (i.e., below a threshold). This relaxation may be convenient from a quantum algorithmic perspective, but it is also still useful from a data-analysis perspective, since

Cheeger’s inequality demonstrates that the magnitudes of the small nonzero eigenvalues of the graph Laplacian characterizes the connectedness of the graph [20], and similar results hold for combinatorial Laplacians [21]. In folklore it is conjectured that for difficult cases, the magnitude of the smallest nonzero eigenvalue of combinatorial Laplacians very often scales inverse polynomially [14], in which case the number of “small” eigenvalues coincides with the number of zero eigenvalues if the threshold is chosen appropriately. While the problem of counting small eigenvalues is more suitable to be solved on a quantum computer, it could turn out to still be  $\text{QMA}_1$ -hard if the TDA matrices have a sufficiently large spectral gap. Specifically, if the TDA matrices are sufficiently gapped, then one could count the number of zero eigenvalues (which is  $\text{QMA}_1$ -hard [17]) by counting the number of eigenvalues below the spectral gap (i.e., the number of “small” eigenvalues).

A related (yet different) problem for which complexity-theoretical results are known is that of estimating *normalized* Betti numbers to within additive inverse polynomial precision. That is, the number of zero eigenvalues divided by the total number of eigenvalues, which here would be  $\beta_{k-1}^G/|\text{Cl}_k(G)|$  (if the TDA matrix is sufficiently gapped). This quantity is natural from a quantum computational complexity perspective (though not from an applications perspective), since a quantum algorithm naturally estimates probabilities (so normalized quantities in this case), and since additive errors allow for a direct relationship to definitions of complexity classes like DQC1.

Specifically, in Ref. [15] it was shown that the generalization of this problem, namely estimating the ratio when allowing a range of small eigenvalues, rather than strictly zero eigenvalues, for *arbitrary* Hermitian operators (i.e., the so-called *low-lying spectral density*) is DQC1-hard. This result was build upon in Ref. [10], where it was shown that the problem remains DQC1-hard when restricting the input to combinatorial Laplacians of general chain complexes. It is unknown whether the hardness persists when further restricting to combinatorial Laplacians of clique complexes, and the closest result to this is the  $\text{QMA}_1$ -hardness result of Ref. [17] for the problem of exact counting. This normalized quantity is not typically studied, and indeed there are concerns that Betti numbers may fail to be large enough to be detectable when normalized (see also Appendix F).

As discussed above, estimates of the normalized Betti number with additive error are more natural from a quantum computational complexity perspective. However, from the perspective of applications, we typically work with (unnormalized) Betti numbers (and perhaps their estimates). For this case, the rescaling from normalized Betti numbers to Betti numbers causes an in general exponential blow up of additive errors, and leads to algorithms, which always have exponential runtimes (for constant error). At the same time, in many applications, we require only small

additive errors when the quantities in question are themselves small. For these reasons here we focus on estimation to within a given relative error; that is, the error in the Betti number divided by the Betti number. That is immune to rescaling and can lead to efficient algorithms in the cases when the Betti numbers are large.

Note that the problem of estimating the low-lying spectral density up to a certain relative error is also **DQC1**-hard. The reason is that the relative error must always be at least as large as the error in the normalized quantity, and estimating the normalized quantity to additive precision  $\epsilon$  is **DQC1**-hard for  $\epsilon = 1/\text{poly}(n)$  [15]. It is unknown whether the hardness result holds for Betti numbers, because they are found by restricting to combinatorial Laplacians, rather than arbitrary Hermitian operators. Generally, the larger the Betti number the more efficient the quantum algorithm will be, which in certain cases results in a polynomial quantum runtime. Examples of cases where the Betti numbers are large are discussed in more detail in Sec. IV.

### III. OPTIMIZATION AND ANALYSIS OF QUANTUM TOPOLOGICAL DATA ANALYSIS

In this section we describe our algorithm in detail. Sections III A, III B, and III C are for preparing a state similar to  $\rho_k^G$  in prior work. That is a combination of states  $|x\rangle$  that correspond to  $k$ -cliques of the graph  $G$ . The general principle is to first prepare a Dicke state, which is explained in Sec. III A. That is an equal superposition of all states with  $k$  ones, of which only a subset will be  $k$ -cliques. Therefore, Sec. III B then describes how to efficiently detect the states out of those that are  $k$ -cliques.

In order to provide a further speedup we then use amplitude amplification, as described in Sec. III C. The key difficulty there is that the number of steps of amplitude amplification depends on the amplitude. We therefore use amplitude estimation separately from the amplitude amplification. This provides a significant advantage over fixed-point amplitude amplification [22], which incurs a logarithmic overhead, because we need only to perform the estimation once but perform the amplification many times. Moreover, our amplitude estimation technique using Kaiser windows is improved over standard amplitude amplification, and can be used in far more general applications.

Then in Sec. III D we describe how to block encode the operator  $B_G$  in order to provide a quantum walk operator that has eigenvalues related to those of  $B_G$ , as in Refs. [23,24]. In particular, we need to find eigenvalues  $\pm 1$  of this walk operator, which correspond to eigenvalue 0 for  $B_G$ . This provides a significant advantage over prior work that was based on simulating a Hamiltonian time evolution under  $B_G$ , because we avoid the overheads inherent in simulating Hamiltonian evolution.

Next, in Sec. III E we show how to use an optimal filter to find eigenstates of  $B_G$  with eigenvalue 0 (or  $\pm 1$  for the walk operator). This method improves over prior work that used phase estimation, which has an overhead due to it providing more information (an estimate rather than just distinguishing between zero and nonzero eigenvalues).

Finally, in Sec. III F we use the amplitude estimation again to estimate the proportion of zero eigenvalues, then provide the overall complexity. The use of amplitude estimation here provides a square-root speedup over work based on classical sampling.

#### A. Generating Dicke states with garbage

In this section, we consider preparing an  $n$ -qubit uniform superposition of Hamming-weight  $k$  basis states (which is allowed to be entangled with garbage states). Such a state is known in the previous literature as the Dicke state. In Ref. [25] it was shown how to prepare a Dicke state with  $\mathcal{O}(nk)$  gates, although these gates included rotations, so there would be a logarithmic factor in the complexity when counting non-Clifford gates.

Because the preparation here allows an entangled state to be prepared between the superposition for the Dicke state and ancilla states, it is possible to prepare the state more efficiently. One approach is to apply a quantum sort to  $n$  registers, then use it to apply an inverse sort to the  $n$  qubits with the first  $k$  set in the state  $|1\rangle$ . This is similar to the approach used for symmetrizing states for chemistry in Ref. [24]. Another approach is to use inequality testing to obtain  $k$  successes. Both those approaches give a factor of  $\log n$  in the number of qubits required, which is costly when  $n$  is large. Throughout we use “log” for base 2, and “ln” for natural logs.

We provide two schemes here. In our first scheme we prepare  $n$  registers with approximately  $\log n$  qubits in equal superposition. This is similar to the first step in Ref. [24] where a sort was used. Here we instead find a threshold such that  $k$  registers are less than or equal to this threshold. This approach is explained in detail in Appendix B 1.

Our second scheme is based on preparing separate superposition states for the positions of each of the individual ones. The target state is obtained by adding all those ones into the target state, but has a higher amplitude for failure arising from ones in the same locations. We provide the details in Appendix B 2. The complexities of these two schemes are as in the following lemma.

*Lemma 2 (Dicke preparation).*—The Dicke state with  $k$  ones in  $n$  qubits may be prepared with probability of success

$$\frac{1}{\binom{cn}{n}} \binom{n}{k} \sum_{\ell=1}^{cn} [\ell^k - (\ell-1)^k] (cn - \ell)^{n-k} \approx 1 - \frac{1}{2c}, \quad (14)$$

using

$$(n_{\text{seed}} + 1) \left[ \frac{n}{2}(n_{\text{seed}} + 2) + \lceil \log n \rceil \right] \quad (15)$$

Toffolis, where  $n_{\text{seed}}$  is a number of seed qubits

$$n_{\text{seed}} := \lceil \log cn \rceil, \quad (16)$$

for some constant  $c$ . Alternatively it may be prepared with probability of success

$$\frac{k!}{n^k} \binom{n}{k} \quad (17)$$

with Toffoli complexity

$$(k + 2)n + k(4\lceil \log n \rceil - 1) + \lceil \log k \rceil, \quad (18)$$

or

$$(k + 2)n - 2k + \lceil \log k \rceil, \quad (19)$$

for  $n$  a power of 2. For this preparation, the state may be entangled with an ancilla system.

Although the first scheme has better asymptotic complexity of  $\tilde{\mathcal{O}}(n)$ , we find that for realistic parameters its complexity is considerably larger. The lower probability of success of the second scheme results in a larger factor in the complexity, so the approach that is optimal will depend on the parameters.

In comparison, prior work in Refs. [9,15] used a procedure based on an integer enumeration of all basis states for the Dicke state. Lloyd *et al.* [8] used a method with a superposition over values of  $k$  that is not directly comparable. The complexity in Ref. [9] does not appear correct (the complexity in Ref. [15] just cites that result). The method it uses is to first compute a Pascal triangle of binomial coefficients up to  $\binom{n}{k}$  with complexity  $\tilde{\mathcal{O}}(n^2k)$ .

To convert a natural number  $l$  to a Hamming-weight  $k$  string, it then starts by finding the largest value of  $x$  such that  $\binom{x}{k} < l$ . It is said that the value of  $x$  can be found using  $\tilde{\mathcal{O}}(k)$  gates via a binary search using Pascal's triangle as a lookup table. The complexity is given as the number of steps in the binary search, which is not correct. The reason is that  $l$  is given in quantum superposition, so  $x$  needs to be searched for in superposition, and finding the appropriate entry in the lookup table (to perform the inequality test for the binary search) has complexity of the size of the lookup table. The value of  $k$  is fixed so not the entire lookup table is needed, but there are  $n$  entries needed, and each has size  $\mathcal{O}(k \log n)$ . This has a complexity of  $\mathcal{O}(kn \log n)$ , which then needs to be performed  $\mathcal{O}(\log n)$  times in the binary search, so would give a complexity  $\mathcal{O}(kn \log^2 n)$ . That complexity needs to be multiplied by  $k$  steps of the

algorithm to give overall complexity  $\mathcal{O}(k^2 n \log^2 n)$  for the conversion.

The complexity of converting in the opposite direction, from a Hamming-weight  $k$  string to a natural number, is given correctly in Ref. [9] as  $\tilde{\mathcal{O}}(nk)$ . The complexity of Pascal's triangle is somewhat less than that given in Ref. [9]. It can be calculated classically and entered into a quantum registers with  $\mathcal{O}(nk \log n)$  Clifford gates, so zero Toffoli complexity. The complexity of preparing an equal superposition over natural numbers is  $\mathcal{O}(k \log n)$ . That is omitted in Ref. [9], which is reasonable because it is smaller than the other complexities.

So in Ref. [9], the leading-order complexity of the Dicke-state preparation is  $\mathcal{O}(k^2 n \log^2 n)$ , which is a factor of  $k^2$  larger than our complexity of  $\mathcal{O}(n \log^2 n)$  for our first approach, and a factor of  $\tilde{\mathcal{O}}(k)$  larger than the complexity of Dicke-state preparation in Ref. [25]. In the example we give in Sec. IV A,  $k = 16$  so the factor of  $k^2$  is 256, and our first approach has about 2 orders of magnitude improvement in the complexity of this step as compared to Ref. [9]. For that example there is about another factor of 4 improvement by using our second approach.

## B. Detecting the cliques

In the previous section, we have discussed the preparation of the  $n$ -qubit Dicke state with Hamming weight  $k$  (and an additional garbage register)

$$\frac{1}{\sqrt{\binom{n}{k}}} \sum_{|x|=k} |x\rangle \frac{1}{\sqrt{k!(n-k)!}} \sum_{\sigma_{(0\dots 01\dots 1)=x}} |\sigma\rangle. \quad (20)$$

Here, the first register holds all  $n$ -qubit strings with Hamming weight  $k$ , representing subsets of  $k$  vertices in an  $n$ -vertex graph  $G$ . We now describe a quantum circuit that detects whether a given string  $x$  represents a  $k$ -clique in the underlying graph, with the promise that  $x$  has Hamming weight  $k$ . Specifically, our goal is to implement the mapping

$$|x\rangle |0\rangle |0\rangle \mapsto |x\rangle |x \in \text{Cl}_k(G)\rangle |\text{garb}_x\rangle. \quad (21)$$

Here, the second register has value 1 if  $x$  represents a  $k$ -clique in  $G$  and 0 otherwise. The third register contains some garbage information  $\text{garb}_x$  that can depend on  $x$  and need not be uncomputed.

Our implementation of the clique detection is related to the approach of Ref. [26]. Specifically, we introduce a register of  $\lfloor \log \binom{k}{2} \rfloor + 1 \leq 2 \log k$  qubits to represent integers  $0, \dots, \binom{k}{2}$ . This register will be used to count the number of edges in the subgraph induced by the  $k$  vertices denoted by  $x$ . For the graph, we assume that it is given by a classical database, so we need to run through this classical database, rather than assuming any oracular access to the graph. Let

us assume that we have a listing of all edges in the graph. That is, for each edge, we have a listing of the two nodes. In order to implement this classical data, for each edge in the list we use a Toffoli with the qubits representing those two nodes as controls, and an ancilla as target. In the case where both qubits are in the state 1, the ancilla qubit will be flipped.

The complexity is then given by a number of Toffolis equal to the number of edges, which we denote  $|E|$ . We aim to sum all the bits output by these Toffolis. Provided that we are restricted to Hamming weight  $k$ , if  $x$  represents a  $k$ -clique then every pair of ones in  $x$  will result in a 1, so the sum will yield  $\binom{k}{2}$ . Summing bits in the obvious way would yield a complexity scaling as  $2|E| \log k$  Toffolis, because each addition requires multiple Toffolis. An improved method is given in Ref. [27], where it would take no more than  $|E|$  Toffolis, but the same number of ancillas would be required, which would typically be a prohibitively large cost. An alternative way of summing bits is given in Ref. [6], where multiple groups of bits are summed, and their sums are summed. The overall complexity is no more than  $2|E|$  Toffolis, and only a logarithmic number of ancillas is used. The costs of the three main parts of the algorithm are as follows.

- (1) There is cost  $|E|$  Toffolis for checking the edges of the graph. The resulting qubits can be erased with measurements and phase corrections, with zero Toffoli cost.
- (2) The complexity of the efficient bit sum approach from Ref. [6] is  $2|E|$ .
- (3) There is complexity no more than  $2 \log k$  Toffolis to check that the output register is equal to  $\binom{k}{2}$ .

Therefore, the total cost of clique detection is no more than  $3|E| + 2 \log k$  Toffolis. In many cases we will need to reflect on the result of this test. In that case, the  $2 \log k$  cost is not doubled, because we can replace the equality test with a controlled phase. Therefore, the cost in that case is  $6|E| + 2 \log k$ . If we were to retain the qubits resulting from the edge checking and use the sum from Ref. [27], the cost would be  $2|E| + 2 \log k$ , though with a large ancilla cost.

Later when we consider the block encoding of the Hamiltonian we will need to allow a wider range of Hamming weights,  $k - 1$ ,  $k$ , and  $k + 1$ , in the case where we are block encoding the Hamiltonian projected onto this subspace. First we can sum the ones in the string  $x$ , which has Toffoli complexity  $n$ . We can check if the sum is equal to  $k - 1$  with  $\lceil \log n \rceil$  Toffolis, then check if it is  $k$  or  $k + 1$  with further Toffolis with the unary iteration procedure. The number of Toffolis needed depends on the value of  $k$ , and some values will require about another  $\lceil \log n \rceil$  Toffolis. For each we can use CNOTs to output a success flag on an ancilla qubit.

We can also output the value of  $\binom{k-1}{2}$ ,  $\binom{k}{2}$ , or  $\binom{k+1}{2}$  in another register. In this case we would need to apply an equality test between the result in our sum register and the result in this register, which again has a Toffoli complexity no larger than  $2 \log k$ . There are also  $n$  Toffolis needed to sum the ones in  $x$  and no more than  $3 \lceil \log n \rceil$  Toffolis to check the number of ones.

Much of the complexity of the algorithm is due to the use of amplitude amplification to find the cliques. There has been much work on quantum algorithms for clique finding, but these algorithms are typically posed in terms of calls to an oracle for the graph, with a possibly large complexity for additional gates. What that means is that the complexity in terms of oracle calls is no more than  $\mathcal{O}(n^2)$  to find all the edges, and then there can be a very large amount of postprocessing to find the cliques.

When there are  $n$  vertices there cannot be any more than  $n(n - 1)/2$  edges, but in practice we would only use the above algorithm for  $|E|$  less than half this. The reason is that for larger numbers of edges, it is more efficient to use a database of *missing* edges. If we use such a database, we can then iterate through all pairs of nodes in the database and use a Toffoli controlled on the corresponding qubits. If we find any cases where this gives one, it means that there is a missing edge and the state does not represent a clique. We therefore need to perform an OR on all the resulting qubits.

Similarly to the case above using the list of edges, one could perform an approach using addition and obtain a complexity with  $|E|$  replaced by  $|E^C|$ , where  $E^C$  is the set of missing edges. But, since we need only to find a single one out of all the results, we can instead use an approach for a multiply controlled Toffoli with a limited number of ancillas. If one is willing to use about  $\sqrt{|E^C|}$  ancillae, then the cost is  $2|E^C|$ , with the same cost for erasure.

In particular, consider grouping the list of missing edges into sets of approximately  $\sqrt{|E^C|}$ . For each we can perform the Toffolis with the qubits representing the corresponding nodes as controls, then perform a multiply controlled Toffoli with those qubits as controls (with the appropriate bit flips to give an OR). The multiply controlled Toffoli has Toffoli cost one less than the number of qubits in the group. So, for example, if  $\sqrt{|E^C|}$  is an integer, then it is cost  $\sqrt{|E^C|} - 1$ . The qubits giving the results of the Toffolis for the missing edges can be erased using Clifford gates similarly as before.

We do this for each of the approximately  $\sqrt{|E^C|}$  groups. Then we perform a multiply controlled Toffoli on the results for all these groups. Because there are about  $\sqrt{|E^C|}$  this is again the Toffoli and ancilla cost. For example, if  $\sqrt{|E^C|}$  is an integer, then there would be Toffoli cost  $(\sqrt{|E^C|} - 1)\sqrt{|E^C|}$  for the  $\sqrt{|E^C|}$  groups, followed by  $\sqrt{|E^C|} - 1$  for the multiply controlled Toffoli on the



results, for cost  $|E^C| - 1$ . That is combined with the  $|E^C|$  cost of the Toffolis for the individual missing edges, for total cost  $2|E^C| - 1$ . Given the improved efficiency of this approach, it would be preferable to using the list of edges for  $|E|$  above about  $n^2/5$ . We can therefore summarize the complexity of the clique checking as follows.

*Lemma 3 (Clique checking).*—It can be checked that an  $n$ -qubit state corresponds to a clique of graph  $G$  using  $3|E| + 2 \log k$  Toffolis given a classical database of edges  $E$ , or  $2|E^C|$  given a classical database of missing edges  $E^C$ . The costs of reflecting on the result of the clique check are  $6|E| + 2 \log k$  or  $4|E^C|$  Toffolis.

The complexity in Ref. [8] is given as  $\mathcal{O}(k^2)$  in terms of calls to an oracle for the distances between the points. Similar complexities are given in Refs. [9,15] but are not explained, so they are likely using the same assumption as Ref. [8]. That complexity is not directly comparable to the result here, because we are instead assuming that we are given an explicit listing of edges (or missing edges).

In order to provide a comparison between that approach and ours, we can consider a slight modification of that where a database of locations of points is given. In that case, one can use a quantum sort on the Dicke state to also sort the locations of the nodes to the first  $k$  data locations. That sort has complexity  $\mathcal{O}(n \log n \log(1/\epsilon))$  given that the locations are given to accuracy  $\epsilon$  (relative to the range of positions). Then the distances can be checked with complexity  $\mathcal{O}(k^2 \log^2(1/\epsilon))$ . The factor of the square of the log here is from the complexity of performing squares for determining the (squared) distance.

We consider an example of a graph in Sec. IV A with  $n = 256$ ,  $k = 16$ , and  $|E| = 30720$ . In that example,  $|E^C| = 1920$ , so it is more efficient to use the list of missing edges in our approach. In comparison, if one were to attempt to use the database of locations, then just the sort would have higher complexity than  $2|E^C|$ . If edges were determined from positions given in a three-dimensional space, then an accuracy of the components of the locations of only 4 bits would result in complexity larger than our costing by an order of magnitude. Although it is difficult to compare our approach to Refs. [8,9,15] due to the different model, we can expect an actual implementation of that type of approach to have at least an order of magnitude larger complexity.

### C. Amplifying the initial state

We aim to amplify the initial state so that we have the state with an equal superposition over cliques. The strategy is to initially estimate the amplitude just once, then apply the appropriate number of steps of amplitude amplification when we are preparing the state to estimate the size of the null eigenspace. It is possible to show that the complexity of estimating the amplitude is as given in the following lemma.

*Lemma 4 (Quantum amplitude estimation).*—Let  $U$  be a unitary and let  $0 < a < 1$  be such that

$$U|0, 0\rangle = a|\psi_0, 0\rangle + \sqrt{1 - a^2}|\psi_1, 1\rangle. \quad (22)$$

There exists a quantum algorithm, which estimates  $a$  to within error  $\epsilon$  with probability of error less than  $\delta$ , using

$$N = \frac{\pi}{\epsilon} \sqrt{1 + \alpha^2} = \frac{1}{2\epsilon} \ln(1/\delta) + \mathcal{O}(\epsilon^{-1} \ln \ln(1/\delta)) \quad (23)$$

calls to  $U$  or  $U^\dagger$ .

The proof for this lemma is given in Appendix D. To see the value of  $\epsilon$  needed, note that the probability of success will be reduced to approximately  $\sin^2((1 \pm \epsilon/a)\pi/2)$  if we incorrectly choose the number of iterates in the amplitude amplification due to imprecision in estimating the amplitude. That translates to a probability of failure of the amplitude amplification of approximately  $(\epsilon\pi/2a)^2$ . For our application, the amplitude is approximately

$$\sqrt{1 - \frac{1}{2c}} \sqrt{\frac{|Cl_k(G)|}{\binom{n}{k}}}, \quad (24)$$

where the first factor comes from failure of the Dicke-state preparation, and the second from the clique checking. For simplicity, in the following expressions for complexity we will omit the factor of  $\sqrt{1 - 1/2c}$ , which is close to 1. The amplitude estimation is needed because it typically will be unknown how many cliques there are  $|Cl_k(G)|$ . Inaccuracy in the amplitude estimation translates to a probability for failure of the amplitude amplification due to using an incorrect number of steps.

In practice the “failure” of the amplitude amplification is not a major problem, because it can be combined into an uncertainty in estimation of the Betti number. That is, in the next step instead of estimating the Betti number relative to  $|Cl_k(G)|$ , we will be estimating it relative to a value that may be increased by about a factor of  $1/[1 - (\epsilon\pi/2a)^2]$  (using the approximation of the sin function). If we want  $(\epsilon\pi/2a)^2$  no more than a relative error  $r$ , then we should choose

$$\epsilon \leq \frac{2\sqrt{r}}{\pi} \sqrt{\frac{|Cl_k(G)|}{\binom{n}{k}}}. \quad (25)$$

That means that the cost would be

$$\frac{\ln(1/\delta)}{\sqrt{r}} \frac{\pi}{4} \sqrt{\frac{\binom{n}{k}}{|Cl_k(G)|}} \quad (26)$$

steps. In comparison, the number of steps of the amplitude amplification is approximately

$$\frac{\pi}{4} \sqrt{\frac{\binom{n}{k}}{|Cl_k(G)|}}. \quad (27)$$

That is, the amplitude estimation is more costly by a factor of  $\ln(1/\delta)/\sqrt{r}$ .

This cost of the Dicke-state preparation from Lemma 2 will be doubled in amplitude estimation and amplification when we account for the need to unprepare the Dicke state. We also need to reflect on the clique check, with complexity  $6|E| + 2 \log k + 4|E^C|$  as described in Lemma 3. The total complexity for each step of the amplitude estimation and amplification is the total of these two complexities.

This approach of separating the estimation and amplification provides a significant improvement over the obvious approach of using fixed-point amplitude amplification [22] to provide amplification with an unknown overlap. That requires a logarithmic factor in the complexity similar to amplitude estimation. In contrast, here we have only that logarithmic factor in the cost once in the initial amplitude estimation, then in the remainder of the algorithm we eliminate the logarithmic factor by just performing amplitude amplification with the initially estimated amplitude.

It is somewhat ambiguous to compare our approach to that in Refs. [8,9,15]. Reference [8] just invokes the “multi-solution version of Grover’s algorithm,” which is not sufficiently specific to give a complexity because there are multiple approaches. Reference [9] cites the version of Grover’s algorithm from Ref. [28], and Ref. [15] just mentions Grover’s algorithm and uses the complexity from Ref. [9]. The problem with citing [28] is that it is not sufficient to specify exactly which approach is intended.

One approach for searching with an unknown number of solutions given in that work is to just use the approach of Ref. [29], which would give a factor of  $9/\pi$  in the complexity, but that approach would not be compatible with a later amplitude estimation used in Ref. [9]. That is because the approach of Ref. [29] relies on a sequence of measurements to obtain success of the search. The measurements would prevent the later amplitude estimation (for the number of zero eigenvalues) being used.

Reference [28] also mentions the approach of performing amplitude estimation, followed by Grover’s algorithm for a known number of solutions. Our proposal here is to divide between using amplitude estimation once, followed by amplitude amplification based on the estimation many times within the rest of the algorithm. That gives a significant improvement over using both at every step (which would be the obvious interpretation of just citing Ref. [28]).

Moreover, we provide a significant improvement in the efficiency of amplitude estimation over that in Ref. [28]. See Theorem 6 of that work for their result in terms of the error in the squared amplitude. Translating that to the error in the amplitude, the number of steps needed is approximately  $\pi/\epsilon$  to obtain  $1 - \delta = 8/\pi^2$ . Repetitions would be needed to obtain a desired  $\delta$ , which would typically be smaller. If, for example,  $\delta = 1/20$  and the number of

repetitions is 5, then our approach gives about an order of magnitude improvement.

#### D. Block encoding the sparse Hamiltonian

Having constructed the sparse oracles in the previous section, we now implement a quantum circuit that block encodes the sparse Hamiltonian. Block encoding is a generalization of a linear combination of unitaries, where an operator  $B$  is given by  $\langle 0|U|0\rangle = B/\lambda$  for a unitary operator  $U$  acting on an ancilla system as well, and  $|0\rangle$  on that ancilla system. Together with a reflection on the ancilla system, it can then be used to construct what was dubbed a “qubitized” or “qubiterate” operator. These principles were introduced in Ref. [23]. We use a similar principle as in Refs. [10,11], except here we are implementing the Dirac operator  $B_G$  rather than the combinatorial Laplacian. In Ref. [10] it is shown that the Dirac operator for *all* Hamming weights and unrestricted by the cliques can be written as

$$B = \sum_{j=1}^n (a_j + a_j^\dagger), \quad (28)$$

where  $a_j$  and  $a_j^\dagger$  are fermionic annihilation and creation operators on qubit  $j$ . Using the usual Jordan-Wigner representation that gives the Hamiltonian

$$\sum_{j=1}^n Z_1 \otimes Z_{j-1} \otimes X_j, \quad (29)$$

where the subscripts indicate the qubits that these operators act on (starting the numbering from 1). This is the core of the implementation of the complete Hamiltonian, and can easily be implemented by first preparing an equal superposition state over  $n$ -basis states, then applying the controlled string of Pauli operators as in Fig. 9 of Ref. [30].

To understand the reason that the Pauli string encodes the matrix, note that  $\partial_k$  will remove a one from some location in the bit string  $x$  of Hamming weight  $k + 1$  and apply a sign according to the number of ones prior to that location. That can be achieved by applying an  $X$  in that location, and applying  $Z$  gates on all qubits prior to that location. We need a superposition of applying the  $X$  in all locations where there are ones. Moreover, we also want to apply  $\partial_{k+1}^\dagger$  to a bit string of Hamming weight  $k + 1$ . This involves flipping a zero to a one (which can be done with an  $X$  gate) and applying a sign according to the number of ones prior to that position. This can again be done using a string of  $Z$  gates. Now we want a superposition of performing  $X$  gates at all locations where there are ones, *and*  $X$  gates where there are zeros, which can be implemented by the above sum of Pauli strings.

Here we aim to block encode the matrix

$$B_G = \begin{bmatrix} 0 & \partial_{k-1}^G & 0 \\ \partial_{k-1}^{G\dagger} & 0 & \partial_k^G \\ 0 & \partial_k^{G\dagger} & 0 \end{bmatrix}. \quad (30)$$

The difference of this from the unrestricted case  $B$  in Ref. [11] is that it acts only on states with Hamming weight  $k-1, k, k+1$ , and gives zero otherwise. Similarly, it gives only states with Hamming weight in this range. Moreover,  $B_G$  is restricted to the clique subspace. That means it must give zero if the input state is not a clique, and must also not give any output states that are not cliques.

Next we provide a general method of constructing a qubiterate operator in cases where tests on the system state are required. The block encoding with the tests can be described as

$$(|0\rangle\langle 0| \otimes P) V (|0\rangle\langle 0| \otimes P) = |0\rangle\langle 0| \otimes B_G/\lambda, \quad (31)$$

where  $P$  is a projection on the system that tests the Hamming weight and cliques. We are adopting notation similar to Eq. (3) in Ref. [24], but replacing the identity with  $P$  to indicate that a projection is needed on the target system. We will assume  $V$  is Hermitian; if it is not we can construct a Hermitian  $V$  by block encoding it as  $V \mapsto V \otimes |1\rangle\langle 0| + V^\dagger \otimes |0\rangle\langle 1|$  [31]. Similarly, we are writing  $B_G$  for the operator we aim to block encode, but this reasoning applies for a more general Hamiltonian  $H$ .

If  $|k\rangle$  is an eigenstate of  $B_G$  with energy  $E_k$  and satisfying  $P|k\rangle = |k\rangle$ , then by definition we must have

$$V|0\rangle|k\rangle = \frac{E_k}{\lambda}|0\rangle|k\rangle + i\sqrt{1 - \left|\frac{E_k}{\lambda}\right|^2}|0k^\perp\rangle, \quad (32)$$

where  $|0k^\perp\rangle$  is defined as a state such that

$$(|0\rangle\langle 0| \otimes P)|0k^\perp\rangle = 0. \quad (33)$$

Then we can define the qubiterate as

$$W := RV, \quad (34)$$

with

$$R := i(2|0\rangle\langle 0| \otimes P - I). \quad (35)$$

This is similar to that in Ref. [24], except we have included the projection  $P$  in the reflection operation. That is, we are applying the tests as part of the reflection, instead of applying them in the operation  $V$ .

Then we obtain

$$W|0\rangle|k\rangle = i\frac{E_k}{\lambda}|0\rangle|k\rangle + \sqrt{1 - \left|\frac{E_k}{\lambda}\right|^2}|0k^\perp\rangle. \quad (36)$$

It is also found that

$$W|\chi k^\perp\rangle = i\frac{E_k}{\lambda}|\chi k^\perp\rangle + \sqrt{1 - \left|\frac{E_k}{\lambda}\right|^2}|\chi\rangle|k\rangle. \quad (37)$$

Here we have corrected a minor error from Ref. [24] where there was an  $i$  appearing on the second term. See Appendix E for the derivation. Then it is easy to see that

$$\frac{1}{\sqrt{2}}(|0\rangle|k\rangle \pm |0k^\perp\rangle) \quad (38)$$

are eigenstates of  $W$  with eigenvalues  $\pm e^{\pm i \arcsin(E_k/\lambda)}$ . This is the usual relation for the eigenvalues of the qubitized operators, showing that this approach for constructing the walk operator works.

For our implementation here,  $V$  is just the controlled string of Pauli operators together with preparation of an equal superposition state. The reflection on the target system expressed by the projector can be implemented by computing an ancilla qubit flagging that the projection is satisfied (we have the appropriate Hamming-weight range and cliques), reflecting on that qubit and the control qubits, then uncomputing the test. In some cases this can give a significant reduction in complexity over performing the test before and after  $V$ . If the ancilla qubits used to compute the tests are retained, then they can be erased with Clifford gates and measurements. For the application here that would be too costly in terms of ancilla qubits, so we incur the Toffoli cost of the test again in erasing the ancillas.

For the complexity of the implementation we have the following costs.

- (1) Preparing an equal superposition state over  $n$ -basis states, which can be performed with complexity  $4\lceil \log n \rceil + 1$  Toffolis [6], or just with Hadamards if  $n$  is a power of 2. This cost is incurred twice.
- (2) The controlled string of Pauli operators can be applied with Toffoli complexity  $n-1$  using the method in Fig. 9 of Ref. [30].
- (3) The Hamming weight can be computed with no more than  $n$  Toffolis and  $n$  ancilla qubits [27]. In that case we would not need to double the complexity for the reflection because the sum can be uncomputed with Cliffords. We could use  $2n$  Toffolis with a logarithmic number of qubits [6], but in that case we would need to double the complexity for the uncomputation cost.
- (4) The complexity of outputting qubits with  $\binom{k-1}{2}$ ,  $\binom{k}{2}$ , or  $\binom{k+1}{2}$  is no more than  $3\lceil \log n \rceil$ . At the same time

we can use the QROM to output a qubit, which flags if the Hamming weight is outside the range. These qubits can be erased with Cliffords by retaining a logarithmic number of ancilla qubits.

- (5) As described above, the cost of the reflection on the clique test is no more than  $6|E| + 2 \log k$  given a database of edges, or  $4|E^C|$  given a database of missing edges, as in Lemma 3.
- (6) Note that there is a reflection on the result of two tests, but this would correspond to a controlled-Z, which is a Clifford gate.

The overall complexity is therefore as in the following lemma.

*Lemma 5 (Block-encoding complexity).*—The Toffoli complexity of block encoding  $B_G/\lambda$  with the operator  $B_G$  as defined in Eq. (6) is

$$6|E| + 2 \log k + 5n + 11 \log n + \mathcal{O}(1), \quad (39)$$

when given a database of edges  $E$ , or

$$4|E^C| + 5n + 11 \log n + \mathcal{O}(1), \quad (40)$$

when given a database of missing edges  $E^C$ . The value of  $\lambda$  for this block encoding is approximately  $n$ .

These complexities come from adding the Toffoli complexities in the list above. The value of  $\lambda$  is obtained by noting that we use a linear combination of  $n$  Pauli strings. This value will be increased very slightly because of imperfect preparation of an equal superposition state in the method of Ref. [6]. That increase is normally less than one part in 1000, so will be ignored here.

In comparison, the approaches in Refs. [9,15] are not very specific about the approach. They give a factor of  $n^2$  for an  $n$ -sparse operator, coming from the general procedure for decomposing an unstructured  $n$ -sparse operator into 1-sparse operators from Ref. [32]. The implementation of the operator also requires checking cliques, which is not addressed in Refs. [9,15].

Ignoring those issues of how the operator is applied, the major difference between the proposal here is that we use a block encoding to construct a walk operator instead of simulating evolution under the Hamiltonian. References [9,15] invoke the results in Refs. [32,33] for the Hamiltonian evolution for unit time. In practice that would need to be adjusted to a time  $1/\lambda$  in the simulation in order to prevent wraparound of the eigenvalues (which would cause nonzero eigenvalues to be measured as zero). For these short times the complexity of the Hamiltonian evolution is multiplied by a logarithmic factor.

If we use the estimate of the complexity from Ref. [33] given in Ref. [34], then we may expect the complexity to be larger than the complexity for the block encoding we have given here by a factor of 6 for the example in

Sec. IV A. There will be more significant factors in other examples with smaller gaps than the example in Sec. IV A.

### E. Projection-based overlap estimation

In order to estimate the number of zero eigenvalues of the Hamiltonian, we project onto the zero eigenspace, then perform amplitude estimation. The projection can be approximated using a Chebyshev polynomial approach. First, recall that in the qubitization the zero eigenvalue of the Hamiltonian is mapped to eigenvalues  $\pm 1$  of the qubitized operator. For the filter function on the phase  $\phi$  of the eigenvalues of the walk operator, one can take

$$\tilde{w}(\phi) = \epsilon T_\ell(\beta \cos(\phi)) \quad (41)$$

for  $\phi$  taking discrete values  $\pi k/\ell$  for  $k$  from  $-\ell$  to  $\ell$ , and where  $\beta = \cosh(\frac{1}{\ell} \cosh^{-1}(1/\epsilon))$ . Taking the discrete Fourier transform of these values gives the window  $w_j$  such that

$$\tilde{w}(\phi) = \sum_{j=-\ell}^{\ell} w_j e^{ij\phi}. \quad (42)$$

Note that  $\tilde{w}(\phi)$  is a function of  $\cos \phi$ , so  $w_j = w_{-j}$ . Moreover, we have values of  $j$  separated by 2. If  $\ell$  is even, then we have even powers of  $\cos \phi$ , and therefore only even  $j$ . This means that it can be regarded as a polynomial in  $e^{2i\phi}$ . We can select between the qubitized walk step and its inverse by controlling on the reflection, so implementing a linear combination of unitaries may be performed with cost  $\ell$ .

The peak for  $\tilde{w}(\phi)$  will be at 0 and  $\pi$ , which is what is needed because the qubitized operator produces duplicate eigenvalues at phases of 0 and  $\pi$ . The width of the operator can be found by noting that the peak is for the argument of the Chebyshev polynomial equal to  $\beta$ , and the width is where the argument is 1, so  $\beta \cos(\phi) = 1$ . This gives us

$$\cosh(\frac{1}{\ell} \cosh^{-1}(1/\epsilon)) \cos(\phi) = 1. \quad (43)$$

The gap in the Hamiltonian is  $\lambda_{\min}$ , which translates to a gap in the qubitized operator of  $\arcsin(\lambda_{\min}/\lambda)$ . Because the width of the peak should be equal to the gap, we can replace  $\phi$  with  $\arcsin(\lambda_{\min}/\lambda)$ , and solving for  $\ell$  gives

$$\ell = \frac{\cosh^{-1}(1/\epsilon)}{\cosh^{-1}(1/\sqrt{1 - (\lambda_{\min}/\lambda)^2})} \leq \frac{\lambda}{\lambda_{\min}} \ln(2/\epsilon). \quad (44)$$

The complexity of the filter on the walk operator may therefore be given as in the following lemma.

*Lemma 6 (Eigenvalue filtering).*—The complexity of filtering out nonzero eigenvalues of  $B_G$  by a factor of  $\epsilon$  is

$$\frac{n}{\lambda_{\min}} \ln(2/\epsilon) \quad (45)$$

calls to the block encoding of  $B_G$ , given that the gap from eigenvalue 0 is at least  $\lambda_{\min}$ .

This lemma is obtained by using  $\lambda = n$  for the block encoding of  $B_G$  in Eq. (44). To determine the appropriate value of  $\epsilon$  to take, note that  $\epsilon$  tells us the multiplying factor for amplitudes for states with eigenvalues outside the gap. The state starts with equal weighting on all eigenvalues, so ideally we should have the amplitude after filtering  $\sqrt{\beta_{k-1}^G/|Cl_k(G)|}$ . If the state amplitudes outside the gap are multiplied by  $\epsilon$ , then the error in the squared amplitude can be at most  $\epsilon^2$ . This corresponds to an error in  $\beta_{k-1}^G$  of  $\epsilon^2|Cl_k(G)|$ , or a relative error of  $\epsilon^2|Cl_k(G)|/\beta_{k-1}^G$ .

In comparison, the approaches in Refs. [8,9,15] are based on phase estimation. They just give the scaling without specifying the method, which is needed to know the constant factors. The best algorithm for phase estimation would correspond to the method we have given in Appendix D. That would have asymptotic complexity similar to the filtering approach given here. To compare the complexities, we first need to note that the accuracy of the phase estimation should be half the gap. This is because if the phase estimate has error half the gap, if the eigenvalue is zero then it could give an estimate of  $\lambda_{\min}/2$ , which could also correspond to an eigenvalue of  $\lambda_{\min}$ . When this factor of 2 is accounted for the phase-estimation approach has similar complexity to the filter. The phase estimation has a somewhat larger complexity in the nonasymptotic regime, though by only about 60%.

That is, there is a moderate improvement over phase estimation even if one were to use the optimal phase estimation introduced in Appendix D. Because Refs. [8,9,15] did not use that method of phase estimation we would provide a larger improvement over those works, though the size of the improvement is ambiguous because they do not specify the method of phase estimation.

## F. Total complexity of algorithm

The Toffoli costs of the algorithm are as follows. In the following we will present the complexity when using the database of edges, then explain the modification for a database of missing edges.

- (1) The preparation of the Dicke state has a leading-order complexity

$$n \log^2 n + \mathcal{O}(n \log n) \quad (46)$$

Toffolis, or approximately  $2kn$  for the two schemes presented in Appendix B. Here we are including a factor of 2 for inversion.

- (2) The cost of checking cliques is given by  $6|E| + 2 \log k$  where we take into account the need to uncompute the result.
- (3) The cost of amplitude estimation is a number of iterations of steps 1 and 2 given as

$$\frac{\ln(1/\delta)}{\sqrt{r}} \frac{\pi}{4} \sqrt{\frac{\binom{n}{k}}{|Cl_k(G)|}}. \quad (47)$$

- (4) The cost of amplitude amplification of the cliques is given by approximately  $(\pi/4)\sqrt{\binom{n}{k}/|Cl_k(G)|}$  of iterations of steps 1 and 2.
- (5) The walk step for the qubitization needs  $6|E| + 5n + 11 \log n + 2 \log k + \mathcal{O}(1)$  Toffolis.
- (6) For the filtering there are  $n/(\lambda_{\min}) \ln(2/\epsilon)$  calls to the block encoding with costs in item 5 above.
- (7) Lastly, we need to perform amplitude estimation on the entire procedure, using approximately  $\log(1/\delta)/2\epsilon$  calls to the amplitude amplification in 4 and filtering in 6.

In comparison, Ref. [9] invoked the amplitude estimation scheme of Ref. [28], which we improve over by about an order of magnitude. Reference [15] just uses classical sampling, which is quadratically more costly. To give the leading-order complexities, the combined cost of steps 1 and 2 is

$$6|E| + n \log^2 n + \mathcal{O}(n \log n). \quad (48)$$

To distinguish the  $\epsilon$ ,  $\delta$ , and  $r$  (relative error) needed in different steps we will use subscripts. The cost of amplitude estimation is then approximately

$$\frac{\ln(1/\delta_1)}{\sqrt{r_1}} \frac{\pi}{4} \sqrt{\frac{\binom{n}{k}}{|Cl_k(G)|}} (6|E| + n \log^2 n). \quad (49)$$

This is expected to be a trivial cost in the overall algorithm, because the amplitude amplification is performed many more times.

For the remainder of the algorithm, we have an initial cost of

$$\frac{\pi}{4} \sqrt{\frac{\binom{n}{k}}{|Cl_k(G)|}} (6|E| + n \log^2 n) \quad (50)$$

for the amplitude amplification for the initial state. Then there is a cost for the block encoding of

$$6|E| + 5n + \mathcal{O}(\log n), \quad (51)$$

for each step. Multiplying by the number of steps needed for filtering, there is a cost

$$\frac{n}{\lambda_{\min}} \ln(2/\epsilon_3)[6|E| + 5n + \mathcal{O}(\log n)]. \quad (52)$$

To determine the appropriate value of  $\epsilon_3$  to take, note that we are measuring a kernel of size  $\beta_{k-1}^G$  as compared to an overall dimension of  $|\text{Cl}_k(G)| \gg \beta_{k-1}^G$ . The relative accuracy in the estimation of  $\beta_{k-1}^G$  will therefore be about  $\epsilon_3^2 |\text{Cl}_k(G)| / \beta_{k-1}^G$  as explained above at the end of Sec. III E. If we aim for relative accuracy  $r_3$ , then we have complexity

$$\frac{n}{2\lambda_{\min}} \ln \left( \frac{4|\text{Cl}_k(G)|}{r_3 \beta_{k-1}^G} \right) (6|E| + 5n + \mathcal{O}(\log n)). \quad (53)$$

Lastly, the amplitude estimation on the entire procedure needs a number of repetitions

$$\frac{\ln(1/\delta_2)}{2\epsilon_2}. \quad (54)$$

But, this amplitude estimation is on a number of steps corresponding to a reflection requiring both the forward and reverse calculations. That introduces a further factor of 2, so we should use

$$\frac{\ln(1/\delta_2)}{\epsilon_2}. \quad (55)$$

Next,  $\epsilon_2$  corresponds to an accuracy of estimating a ratio  $\sqrt{\beta_{k-1}^G / |\text{Cl}_k(G)|}$ . If we want a relative accuracy  $r_2$ , then the error-propagation formula gives

$$\begin{aligned} r_2 &= \frac{\Delta \beta_{k-1}^G}{\beta_{k-1}^G} = \frac{\epsilon_2}{\beta_{k-1}^G} \left( \frac{d}{d\beta_{k-1}^G} \sqrt{\frac{\beta_{k-1}^G}{|\text{Cl}_k(G)|}} \right)^{-1} \\ &= 2\epsilon_2 \sqrt{\frac{|\text{Cl}_k(G)|}{\beta_{k-1}^G}}, \end{aligned} \quad (56)$$

where  $\Delta \beta_{k-1}^G$  is uncertainty in  $\beta_{k-1}^G$ . In terms of  $r_2$ , the number of repetitions becomes

$$2 \frac{\ln(1/\delta_2)}{r_2} \sqrt{\frac{|\text{Cl}_k(G)|}{\beta_{k-1}^G}}. \quad (57)$$

Applying this to the complexity required for each step, we get a complexity of approximately

$$\begin{aligned} &\frac{\ln(1/\delta_2)}{r_2} \sqrt{\frac{|\text{Cl}_k(G)|}{\beta_{k-1}^G}} \left[ \frac{\pi}{2} \sqrt{\frac{\binom{n}{k}}{|\text{Cl}_k(G)|}} (6|E| + n \log^2 n) \right. \\ &\left. + \frac{n}{\lambda_{\min}} \ln \left( \frac{4|\text{Cl}_k(G)|}{r_3 \beta_{k-1}^G} \right) (6|E| + 5n) \right]. \end{aligned} \quad (58)$$

This is the expression given as Eq. (12) in Lemma 1. If we use the second Dicke preparation scheme, then  $n \log^2 n$  would be replaced with  $2kn$ , and  $\binom{n}{k}$  replaced with  $n^k/k!$ , which gives the expression in Eq. (13) of Lemma 1. For this complexity the factor of  $6|E|$  at the beginning is for the complexity from the database of edges; for the database of missing edges this factor would be replaced with  $4|E^C|$ .

Comparing this to the amplitude estimation cost in Eq. (49) the primary difference is the factor of

$$\sqrt{\frac{|\text{Cl}_k(G)|}{\beta_{k-1}^G}} \quad (59)$$

here. There is another difference in that the amplitude estimation cost has the factor  $1/\sqrt{r_1}$  rather than  $1/r_2$ , so the scaling in terms of relative error is improved. In cases where the number of cliques  $|\text{Cl}_k(G)|$  is much larger than the Betti number  $\beta_{k-1}^G$ , then the amplitude estimation cost is trivial.

We will have a total probability of failure  $\delta = \delta_1 + \delta_2$  due to the two amplitude estimations, and a total relative error  $r_1 + r_2 + r_3$ . In order to reduce the complexity, we can use the fact that the cost of the initial amplitude estimation is much smaller, so we can take  $\delta_1$  and  $r_1$  smaller without much impact on the overall complexity. The  $r_3$  appears inside a logarithm, so can be taken to be smaller than  $r_3$ .

To give the scaling of the complexity in a simpler way, we can simply ignore the amplitude estimation complexity, and replace  $\delta_2$  with  $\delta$ , and replace both  $r_2$  and  $r_3$  with  $r$  (since  $r_3$  can be taken as for example  $r/20$  without much impact on the overall complexity). We will also omit terms of complexity  $kn$  or  $n$  as compared to  $|E|$ . That then gives

$$\begin{aligned} T(G, k, r, \delta) &:= 6|E| \frac{\ln(1/\delta)}{r} \sqrt{\frac{|\text{Cl}_k(G)|}{\beta_{k-1}^G}} \left[ \frac{\pi}{2} \sqrt{\frac{\binom{n}{k}}{|\text{Cl}_k(G)|}} \right. \\ &\left. + \frac{n}{\lambda_{\min}} \ln \left( \frac{4|\text{Cl}_k(G)|}{r \beta_{k-1}^G} \right) \right], \end{aligned} \quad (60)$$

where  $T(G, k, r, \delta)$  gives the required number of Toffoli gates to estimate, with precision parameters  $r, \delta$ , the  $(k-1)$ th order Betti number of a graph  $G$  with  $n$  nodes,  $|E|$  edges and a Laplacian with gap  $\lambda_{\min}$ .

This expression for the complexity is in terms of the relative accuracy  $r$ . Alternatively, if we aimed for a given absolute accuracy  $\alpha$ , then  $\alpha = r \beta_{k-1}^G$ , and so the expression

for the complexity becomes

$$T(G, k, r, \delta) := 6|E| \frac{\ln(1/\delta)}{\alpha} \sqrt{|\text{Cl}_k(G)| \beta_{k-1}^G} \left[ \frac{\pi}{2} \sqrt{\frac{\binom{n}{k}}{|\text{Cl}_k(G)|}} + \frac{n}{\lambda_{\min}} \ln \left( \frac{4|\text{Cl}_k(G)|}{\alpha} \right) \right]. \quad (61)$$

We now discuss the complexity of just the first term in the square brackets, which corresponds to the state-preparation cost rather than the filtering cost. That cost will be dominant if the gap is large, though it must be emphasized that the gap will be small in many cases. This first term for the cost gives

$$T(G, k, r, \delta) = 3\pi |E| \frac{\ln(1/\delta)}{r} \sqrt{\frac{\binom{n}{k}}{\beta_{k-1}^G}}. \quad (62)$$

If we are aiming for a given absolute accuracy  $\alpha$  in  $\beta_{k-1}^G$ , then the complexity would be

$$T(G, k, r, \delta) = 3\pi |E| \frac{\ln(1/\delta)}{\alpha} \sqrt{\binom{n}{k} \beta_{k-1}^G}. \quad (63)$$

The complexity is now larger for large Betti number  $\beta_{k-1}^G$ . The reason for this is that the amplitude estimation is estimating the *square root* of  $\beta_{k-1}^G$ . The square root has a small derivative for large values of  $\beta_{k-1}^G$ , making it more difficult to estimate the Betti number with small absolute error. Again, note that the last three expressions above are only for the state-preparation cost, without the filtering cost.

To compare to the complexity of classical approaches, an exact diagonalization approach would tend to scale as  $\binom{n}{k}^2$ , whereas approximate schemes scale as  $\binom{n}{k}$ . Thus the quantum algorithm would give approximately a square-root speedup over these classical algorithms if  $\beta_{k-1}^G$  is on the order of a constant and one is targeting a fixed relative error estimate. On the other hand, for graphs with large  $\beta_{k-1}^G$ , a speedup that is greater than a square root can be obtained for fixed relative error estimates.

#### IV. REGIMES FOR QUANTUM SPEEDUP

In this section, we ask if there exist regimes where our quantum algorithm offers a significant speedup over the best classical algorithms. The aim is to compute to relative error the  $(k-1)$ th Betti number of the clique complex of a graph  $G$ . Say  $G$  has  $n$  nodes,  $|E|$  edges,  $r$  is the desired multiplicative error, and  $\lambda_{\min}$  is the spectral gap of the combinatorial Laplacian  $\Delta_{k-1}^G = \partial_{k-1}^{G\dagger} \partial_{k-1}^G + \partial_k^G \partial_k^{G\dagger}$ . To simplify the arguments, we will represent the quantum

complexity of this problem as

$$T_q = \tilde{\mathcal{O}} \left( \frac{n|E|}{r \lambda_{\min}} \sqrt{\frac{1}{\beta_{k-1}} \binom{n}{k}} \right). \quad (64)$$

Comparing to Eq. (60), this will asymptotically upper bound both terms up to log factors.

For a rough estimate of the cost of computing the Betti number classically, one could use  $|\text{Cl}_k(G)|$  (i.e., the number of  $k$ -cliques) or  $\binom{n}{k}$ . The reason is that classical algorithms typically start by constructing a list of  $k$ -cliques, and afterwards compute the nullity of the combinatorial Laplacian or boundary operator. The cost of this second step (i.e., estimating the nullity of the combinatorial Laplacian or boundary operator), scales at best linearly in size of the matrix  $|\text{Cl}_k(G)|$  [35]. On the other hand, the first step (i.e., listing all  $k$ -cliques) can be done using a brute force search at cost  $\binom{n}{k}$ . There are more efficient algorithms for listing cliques, though the complexity tends to be dependent on the properties of the graph. However,  $|\text{Cl}_k(G)|$  always lower bounds the cost of listing all the  $k$ -cliques. Therefore,  $|\text{Cl}_k(G)|$  and  $\binom{n}{k}$  can be considered to be lower and upper bounds on the scaling of the classical complexity, respectively. In conclusion, the best classical algorithms for this problem have scaling lower bounded by

$$T_c = \Omega(|\text{Cl}_k(G)|). \quad (65)$$

Recall  $\text{Cl}_k(G)$  is the set of cliques of size  $k$ , which form the  $(k-1)$ -simplices of the simplicial complex. Classical algorithms have extra factors in the complexity, such as  $1/r^2$  dependence on the required precision, that introduce orders of magnitude over this lower bound. Another category of classical algorithms to compare to are those tailored for the specific regime where quantum algorithms are most efficient. Notable examples include the algorithm developed in Sec. IV D, and the algorithm of Apers *et al.* [36]. We defer their comparison to Sec. IV D, where we will highlight regimes where the examples introduced in Sec. IV A continue to exhibit a superpolynomial speedup.

The quantum algorithm will offer a speedup on instances where  $\beta_{k-1}$  is large, and where  $\lambda_{\min}$  is not too small. We can remove the dependence on  $\lambda_{\min}$  if we instead focus on computing an approximate Betti number, in the following sense.

*Definition 1.*—The  $\delta$ -approximate  $k$ th Betti number is  $B_k^\delta = \dim \{v \in \mathcal{H}_{k-1}^G : v^\dagger \Delta_k v / v^\dagger v \leq \delta\}$ . Note  $B_k^0 = \beta_k$ , and in general  $B_k^\delta \geq \beta_k$ .

The same quantum algorithm computes  $B_{k-1}^\delta$  to relative error with cost

$$T_q = \tilde{\mathcal{O}} \left( \frac{n|E|}{r \delta} \sqrt{\frac{1}{B_{k-1}^\delta} \binom{n}{k}} \right). \quad (66)$$

### A. A family of graphs with large Betti numbers and large spectral gaps

In this section, we will construct a family of graphs with all the necessary parameters to enable a large quantum speedup. Our objective here is to demonstrate the existence of instances that fulfill all the prerequisites for the quantum algorithm to achieve a superpolynomial quantum speedup.

Let  $K(m, k)$  be the  $k$ -partite complete graph, where each partition contains  $m$  vertices. That is,  $K(m, k)$  consists of  $k$  clusters, each with  $m$  vertices; there are no edges within clusters, but all edges between clusters are included. Note  $K(m, 1)$  is a collection of  $m$  points with no edges.  $K(m, k)$  gives a useful example of a clique complex with a high Betti number [37]. It also has a Laplacian with a large spectral gap. An example is provided in Fig. 1.

*Proposition 1.*—The  $(k - 1)$ th Betti number of (the clique complex of)  $K(m, k)$  is

$$\beta_{k-1} = (m - 1)^k. \quad (67)$$

*Proposition 2.*—The combinatorial Laplacian  $\Delta_{k-1} = \partial_{k-1}^\dagger \partial_{k-1} + \partial_k \partial_k^\dagger$  of (the clique complex of)  $K(m, k)$  has spectral gap

$$\lambda_{\min} = m. \quad (68)$$

We prove these in Appendix F using techniques from simplicial homology. A further useful fact is that

$$|\text{Cl}_k(K(m, k))| = m^k. \quad (69)$$

Standard classical approaches need to at least store a vector of this length, so we can give a classical complexity as

$$T_c \sim e^{k \ln m}. \quad (70)$$

As a first approximation for the quantum cost, we use the formula in Eq. (62) and consider just the square-root factor and  $|E|$ . In fact, for this example the large number of edges means that it is better to use the list of missing edges to

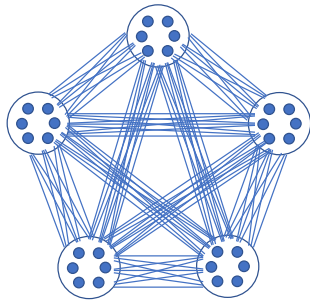


FIG. 1. The graph  $K(6, 5)$ . We do not explicitly provide all the edges from points in a cluster to all other points in other clusters, to avoid overcrowding the figure.

give complexity proportional to  $|E^C|$ . Bearing in mind that  $n = mk$ , Stirling's approximation gives

$$\begin{aligned} \binom{n}{k} &\simeq \frac{1}{\sqrt{2\pi}} \sqrt{\frac{mk}{(n-k)k}} \frac{(mk)^n}{(n-k)^{n-k} k^k} \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{m}{n - \frac{n}{m}}} \left( \frac{m}{(m-1)^{1-\frac{1}{m}}} \right)^n \\ &\leq \left( \frac{m}{(m-1)^{1-\frac{1}{m}}} \right)^n, \end{aligned} \quad (71)$$

where in the first line we have omitted the exponentials in Stirling's approximation because they cancel. Proposition 1 gives  $\beta_{k-1} = (m - 1)^{n/m}$ , giving a quantum complexity scaling as

$$T_q \sim |E^C| \frac{n}{m} \left( \frac{m}{m-1} \right)^{n/2} \leq n^2 e^{k/2}. \quad (72)$$

Therefore, for constant  $m$ , there is a polynomial speedup by a  $2 \ln m$  root (ignoring the  $n^2$  factor). Alternatively, taking  $k$  constant, the above formulae give

$$T_c = \mathcal{O}(n^k), \quad (73)$$

$$T_q = \mathcal{O}(n^2). \quad (74)$$

Then there is a polynomial speedup by a  $k/2$  root. To obtain a superpolynomial speedup,  $m$  can be taken to increase close to linear in  $n$ , but  $k$  can be taken to also increase with  $n$ . Close to the best result is obtained for  $k = c \ln^2 n$  with some constant  $c$ . Then the logs of the complexities are approximately

$$\ln T_c \sim c \ln^3 n, \quad (75)$$

$$\ln T_q \sim 2 \ln n + (c/2) \ln^2 n. \quad (76)$$

That implies a speedup by a  $2 \ln n$  root, which is superpolynomial.

This is still not an exponential speedup, but as far as the graph is concerned this is the best speedup that could be obtained from this type of approach. This is because, with  $k$  constant, the quantum complexity ignoring the  $|E^C|$  factor is  $\mathcal{O}(1)$ . The Betti number is already scaling the same as  $\binom{n}{k}$ , but the overhead from  $|E^C|$  means that the speedup is not exponential.

Next we provide numerical results for the Toffoli complexity as a function of  $n$  and  $k$ . For these calculations we have made a number of adjustments to our simplified expressions in order to provide more accurate results. In particular, we compute the integral of the Kaiser window rather than using the asymptotic expression, as well



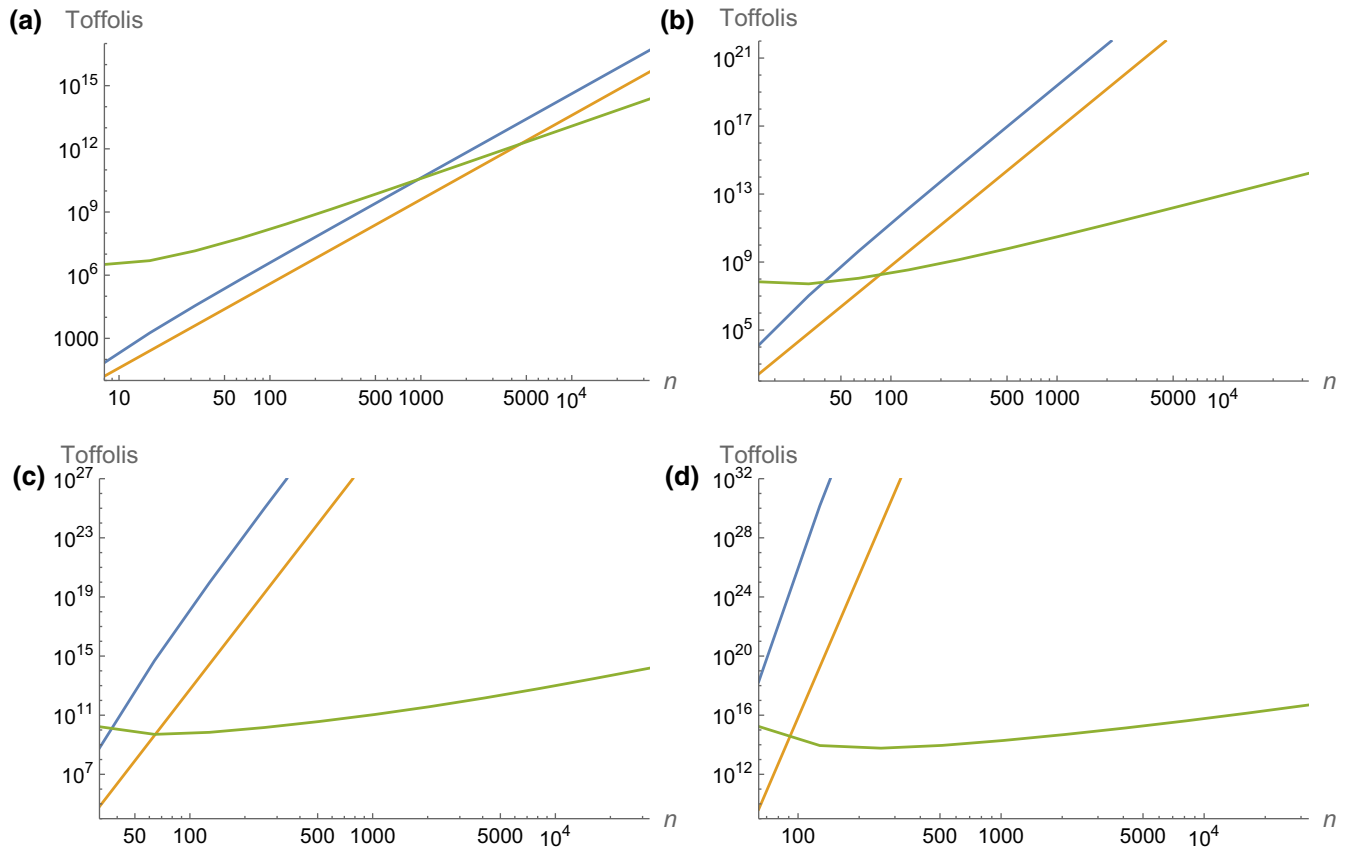


FIG. 2. The Toffoli counts for the quantum algorithm for the Betti number of  $K(m, k)$  as a function of  $n$  for  $k = 4$  (a),  $k = 8$  (b),  $k = 16$  (c), and  $k = 32$  (d). The lines are green for the quantum complexity, blue for  $\binom{n}{k}$ , and orange for the number of cliques  $m^k$ . The values of  $\delta$  and  $r$  are held constant at  $1/20$ . For the relative precision required  $r$ , the value of  $r_2$  (filtering error) is taken to be  $r/20$ , and  $r_3$  (the amplitude estimation error) is taken to be  $r \times 0.95$ .

as including the Dicke state-preparation cost and the initial amplitude estimation cost for the number of steps needed for the state preparation. We are also using the second Dicke preparation scheme from Appendix B, which provides a smaller complexity for this example.

The results are as given in Fig. 2 as a function of  $n$  for a range of values of  $k$ . It can be seen that the cost of the quantum algorithm for a given  $k$  scales approximately as  $n^2$ , with the cost scaling primarily coming from the number of edges in the graph. The classical cost given as the number of  $k$ -cliques or  $\binom{n}{k}$  has similar scaling, which is considerably worse than for the quantum algorithm, and is much worse for larger values of  $k$ , as expected from the analysis above.

For the example of  $n = 256, k = 16$  the quantum cost is approximately 6.8 billion Toffolis, which is comparable to gate counts for classically intractable instances of quantum chemistry [38]. In contrast, the number of cliques is about  $2 \times 10^{19}$ , and  $\binom{n}{k} \approx 10^{25}$ . These numbers are sufficiently large that it should be classically intractable for any method that scales as  $|\text{Cl}_k(G)|$ . For example, just storing the vector would be beyond the storage capacity

of supercomputers. Potentially, more advanced classical algorithms that do not need to store the vector could be tractable.

To compare to the scheme as presented in Refs. [9,15], we improve by about 2 orders of magnitude for this example by using a more efficient Dicke state-preparation scheme. We have a further order of magnitude improvement in complexity by using optimal quantum amplitude estimation in the final step. That gives at least 3 orders of magnitude improvement, which is the difference between a quantum computer running for a day versus years. The total improvement is unclear because some parts of the algorithm were not specified in Refs. [9,15].

We obtain about another order of magnitude improvement by separately performing an amplitude estimation to avoid needing to repeatedly perform it in the initial state preparation. The question of how this would be performed was not addressed in Refs. [9,15]. There is a more modest improvement in using the optimal filter as compared to optimal phase estimation. But, the optimal phase estimation is a procedure introduced here, and there would be larger improvement over less efficient phase estimation.

The type of phase estimation was not addressed in prior work. We also provide an improved clique-checking procedure, but Refs. [9,15] gave complexities in terms of oracle calls without addressing how they would be implemented in terms of elementary gates, making a direct comparison of complexities impossible.

### B. Erdős-Rényi graphs

The family of graphs in Sec. IV A is specifically constructed to have high Betti number and large spectral gap. One might wonder what speedups are *generically* possible. To shed light on this question, we examine the Erdős-Rényi family of random graphs.

The Erdős-Rényi random graph  $G(n, p)$  has  $n$  vertices, and each of the  $\binom{n}{2}$  edges is present i.i.d. with probability  $p$ . In Ref. [39], the following theorem is established.

*Theorem 1.*—Let  $p = n^\alpha$ . If  $-1/k < \alpha < -1/(k+1)$ , then

$$\frac{\beta_k}{\binom{n}{k+1} p^{\binom{k+1}{2}}} \rightarrow 1 \quad \text{almost surely} \quad (77)$$

On the other hand, if  $\alpha < -1/k$  or  $\alpha > -1/(2k+1)$ , then  $\beta_k \rightarrow 0$  almost surely.

Taking  $p = n^{-1/(k+1/2)}$  gives  $\beta_k \sim \binom{n}{k+1} n^{-k/2}$  almost surely. Ignoring the factor  $n|E|/\lambda_{\min}$ , our quantum algorithm can compute the  $k$ th Betti number in time scaling as  $T_q \sim n^{k/4+2}$  for constant  $k$ . For large  $k$ , where the +2 coming from  $|E^C|$  is negligible, this is approximately a quartic speedup.

### C. Rips complexes

One of the main applications of topological data analysis is to Rips complexes induced by finite-dimensional data in  $\mathbb{R}^d$ . This is another shortcoming of the graph family from Sec. IV A—they are defined as abstract graphs, rather than being induced from finite-dimensional data. But are such speedups possible for Rips complexes? Unfortunately, there are results which prevent these large speedups.

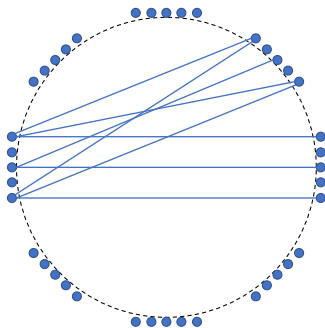


FIG. 3. The Rips complex  $\mathcal{R}_1(S)$ .

It is shown in Ref. [40] that, for any fixed  $k$  and  $d$

$$\max_{S \subset \mathbb{R}^d: |S|=n} \frac{\beta_k(\mathcal{R}_\epsilon(S))}{n^k} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (78)$$

In Ref. [41], the author studies a setting where  $n$  data points are drawn from a fixed underlying probability measure on  $\mathbb{R}^d$ . This is arguably the setting of interest in topological data analysis. They show that the Betti numbers of the derived Rips complexes have three “phases” depending on the scale  $\epsilon$ . (Recall that we include an edge if two points are within distance  $\epsilon$ .) For small  $\epsilon = o(n^{-1/d})$ , called the subcritical phase, the Betti numbers vanish asymptotically. Intuitively the complex is highly disconnected, since we are below the percolation threshold. There is a critical phase  $\epsilon \sim n^{-1/d}$ , where the Betti numbers will scale linearly  $\beta_k \sim n$ . Then for large  $\epsilon = \omega(n^{-1/d})$ , in the supercritical phase, the Betti number grows sublinearly  $\beta_k = o(n)$ . Thus in all regimes, the Betti number grows at most linearly in the number of points. This is of course far from the  $n^k$  scaling needed for superpolynomial speedup.

However, it is possible to construct a Rips complex with large Betti number and large spectral gap, even in  $\mathbb{R}^2$  [40]. We illustrate an example of such a complex  $\mathcal{R}_1(S)$  in Fig. 3.

Construct  $S \subset \mathbb{R}^2$  as follows. Let  $m = n/2k$ ,  $\theta = \pi/k$ , and  $\delta = n^{-4}$ . For  $i = 1, \dots, k$ , let  $x_i^+ = (1/2, i\delta)$  and  $x_i^- = (-1/2, i\delta)$ . Let  $S_0 = \{x_1^+, \dots, x_m^+, x_1^-, \dots, x_m^-\}$ . For  $j = 1, \dots, k-1$ , construct  $S_j$  by rotating  $S_0$  about the origin by an angle  $j\theta$ . Then finally  $S = S_0 \cup \dots \cup S_{k-1}$ . We will take the Rips complex  $\mathcal{R}_1(S)$  with  $\epsilon = 1$ .  $\{x_1^+, \dots, x_m^+\}$  and  $\{x_1^-, \dots, x_m^-\}$  become  $m$ -simplices. There is an edge  $(x_i^+, x_i^-)$  for every  $i$ , but no edges  $(x_i^+, x_j^-)$  for  $i \neq j$ . Due to the small value of  $\delta$ , each  $S_i$  is completely connected to every other  $S_j$ .

*Proposition 3.*—The  $(2k-1)$ th Betti number of  $\mathcal{R}_1(S)$  is

$$\beta_{2k-1}(S) = (m-1)^k = \left(\frac{n}{2k} - 1\right)^k. \quad (79)$$

*Proposition 4.*—The combinatorial Laplacian  $\Delta_k = \partial_k^\dagger \partial_k + \partial_{k+1} \partial_{k+1}^\dagger$  of  $\mathcal{R}_1(S)$  has constant spectral gap  $\lambda_{\min}$ .

We prove these in Appendix F using techniques from simplicial homology. Our quantum algorithm can compute the  $k$ th Betti number in time scaling as  $T_q \sim n^{3+k/4}$  for constant  $k$ .

### D. Randomized classical algorithms for Betti-number estimation

While the previous discussion shows there are cases where quantum algorithms can provide superpolynomial advantages with respect to deterministic classical algorithms for TDA, the question of how randomized classical algorithms perform in this setting is comparably

understudied. There are works studying generalizations of random-walk operators corresponding to higher-order Laplacians for simplicial complexes [42–44]. These are specific to the combinatorial Laplacian context and while they could lead to more efficient classical approaches, no such result is known at the present.

Here we show, perhaps surprisingly, that there exists a randomized classical algorithm, which can compute normalized Betti numbers in the clique dense case using a polynomial number of operations under appropriate assumptions. This shows that the sufficient conditions needed for quantum algorithms to provide an advantage are more subtle than anticipated and that simply having a high-dimensional vector space does not necessarily guarantee a superpolynomial speedup.

The main idea of our algorithm is to use imaginary-time evolution to create a projector onto the kernel of  $\Delta_{k-1}^G = \partial_{k-1}^{G\dagger} \partial_{k-1}^G + \partial_k^G \partial_k^{G\dagger}$ . More specifically, we focus on the Dirac operator  $B_G$  and look at simulating its imaginary-time dynamics of its square using path integral Monte Carlo. We further simplify this approach by taking  $\tilde{B}_G$  to be an analogous operator to  $B_G$  except we now use an energy penalty to penalize any configuration that is not a clique or of the correct parity. In particular,

$$\tilde{B}_G^2 = B_G^2 + \gamma_{\min}(1 - P), \quad (80)$$

where  $P$  as before is the projector onto the states of appropriate Hamming weight and configurations that correspond to a clique, and  $\gamma_{\min}$  is an upper bound on the spectral gap of  $B_G^2$ , which coincides with the second smallest eigenvalue of the combinatorial Laplacian. Further, it is easy to see that if a vector is in the kernel of  $\tilde{B}_G^2$  it is also in the kernel of  $B_G$ . Following the same reasoning as before, as  $B_G$  is Hermitian so is  $\tilde{B}_G$  and thus it has a complete set of orthonormal eigenvectors. This implies that any unit vector  $|\psi\rangle$ , which is supported on  $\mathcal{H}_k^G$  can be decomposed as

$$|\psi\rangle = \cos(\theta) |\psi_g\rangle + \sin(\theta) |\psi_b\rangle, \quad (81)$$

where  $|\psi_g\rangle$  is the projection of  $|\psi\rangle$  onto the kernel of  $\tilde{B}_G^2$  and  $|\psi_b\rangle$  is its orthogonal complement. Then

$$e^{-\tilde{B}_G^2 t} |\psi\rangle = \cos(\theta) |\psi_g\rangle + e^{-\tilde{B}_G^2 t} \sin(\theta) |\psi_b\rangle. \quad (82)$$

Let  $\tilde{B}_G^2 |\lambda_\mu\rangle = \lambda_\mu |\lambda_\mu\rangle$  such that  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{d_{k-1}}$ , where  $d_{k-1} = \binom{n}{k}$ . The operator  $\tilde{B}_G^2$  is positive semidefinite and thus

$$\begin{aligned} \langle \psi | e^{-\tilde{B}_G^2 t} | \psi \rangle &= \cos^2(\theta) + \sin^2(\theta) \langle \psi_b | e^{-\Delta_{k-1}^G t} | \psi_b \rangle \\ &\leq \cos^2(\theta) + \sin^2(\theta) e^{-\gamma_{\min} t}. \end{aligned} \quad (83)$$

If we then pick

$$t \geq \log(1/\epsilon) / \gamma_{\min}, \quad (84)$$

where  $\gamma_{\min}$  is the smallest nonzero eigenvalue of  $\tilde{B}_G^2$ , the expectation value will be at most  $\cos^2(\theta) + \mathcal{O}(\epsilon)$ .

If  $|\psi\rangle$  is chosen such that it is a column of a Haar random unitary over the constrained parity subspace  $\mathcal{H}_k^G$ , the expectation value will be

$$\mathbb{E}_{\text{Haar}}(\cos^2(\theta)) = \frac{\beta_{k-1}}{d_{k-1}}. \quad (85)$$

Thus performing imaginary-time evolution and a Haar expectation value will give the required normalized Betti number.

The remaining question centers around whether the imaginary-time evolution can be performed on a classical computer in polynomial time. First, let us consider a decomposition of the Hamiltonian of the form

$$\tilde{B}_G^2 = \sum_{p=1}^D c_p H_p, \quad (86)$$

where each  $H_p$  is one sparse, Hermitian, and unitary. Hence the eigenvalues of each are  $\lambda_{p_i, v_i} = \pm c_p$ , where  $v_i$  is an index of the eigenvalue and  $p_i$  is the index of the Hamiltonian [45,46]. The Jordan-Wigner decomposition on  $B$  provides such a decomposition and the projector  $P$  can always be written as a sum of a reflection over computational basis states and an identity gate, which provides an efficient decomposition into one-sparse Hermitian and unitary terms.

With this decomposition in hand, we focus on using a path-integral Monte-Carlo simulation of  $\exp(-\tilde{B}_G^2 t)$ . The path-integral expansion works by first breaking up  $e^{-\tilde{B}_G^2 t}$  into  $r$  timeslices, Trotterizing over the matrices in our one-sparse Hermitian decomposition of  $\tilde{B}_G^2$  (which is efficient to determine [45,46]), and then expanding each one-sparse Hermitian matrix in its eigenbasis. Since one-sparse Hermitian matrices can be efficiently diagonalized, this process is classically efficient. (Hermitian one-sparse matrices can be decomposed as a direct sum of one- and two-dimensional matrices, which are trivial to diagonalize.)

Let  $\Gamma$  denote a particular path of eigenvectors in the path-integral representation,  $\mathcal{W}(\Gamma)$  be the product of overlaps between the eigenvectors and  $\lambda_{p_i, \Gamma_i}$  be the eigenvalue corresponding to the eigenvector that appears in the  $i$ th step in the path  $\Gamma$ . Finally, let  $\text{Pr}(\Gamma)$  be a probability distribution from which the paths are drawn that can be chosen to reduce the variance (as is standard in importance sampling). We show in Appendix G that taking the Haar-expectation of the result leads to

$$\mathbb{E}_{\text{Haar}}(\cos^2(\theta)) = \frac{1}{d_{k-1}} \mathbb{E}_{\Gamma} \left( \frac{\exp\left(-\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/2r\right) \mathcal{W}(\Gamma) \delta_{k_1, k_{2rD}}}{\Pr(\Gamma)} \right). \quad (87)$$

We then average over a finite ensemble of these random paths to estimate the expectation value drawn from an appropriate probability distribution. We propose, for general purposes, a Metropolis-Hastings-based algorithm for selecting appropriate paths in the decomposition that are unlikely to have zero values of  $\mathcal{W}$ . More specifically, the algorithm works by drawing an initial eigenstate of the first term in the one-sparse decomposition of the Dirac operator  $B_G$  uniformly. Then a path is drawn by transitioning to one of the two possible connected eigenstates for it randomly. As the terms are Hermitian by assumption, all eigenvalues at each step in the path integral are the same up to a sign. This means that there are only two choices when constructing a path: either we choose to traverse the positive eigenvalue or the negative eigenvalue. Thus each path can be described using  $\mathcal{O}(\log(d_{k-1})rD)$  bits. A path that has nonzero overlaps between the neighboring eigenstates can then be selected in  $\mathcal{O}(\log(d_{k-1})rD)$  time. This is used as an initial guess that is improved using Metropolis-Hastings, wherein the probability of transitioning between two randomly chosen paths  $\Gamma^{(a)}$  and  $\Gamma^{(b)}$  is

$$P(\Gamma^{(b)} | \Gamma^{(a)}) = \frac{\exp\left(-2\lambda_{p_1, \Gamma_1^{(b)}} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i^{(b)}} t/r\right)}{\exp\left(-2\lambda_{p_1, \Gamma_1^{(a)}} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i^{(a)}} t/r\right)}. \quad (88)$$

The equilibrium distribution leads to a thermal distribution over the path  $K$  with

$$\Pr(\Gamma) = \frac{\exp\left(-2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r\right) \delta_{\Gamma \in S_{\Gamma}}}{\sum_{\Gamma \in S_{\Gamma}} \exp\left(-2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r\right)}, \quad (89)$$

where  $S_{\Gamma}$  is the set of all paths. The number of such updates needed to achieve this distribution (within fixed error) scales as  $\mathcal{O}(1/\gamma_M)$  where  $\gamma_M$  is the gap of the Markov chain. We show in Appendix G that, provided the gap is large, this distribution can be efficiently sampled from and forms a good choice for the importance distribution for the paths that minimizes the variance over the paths.

We ultimately find that the number of arithmetic operations needed to estimate the ratio of the kernel to the size of the set of all  $k$ -simplices within additive error  $\epsilon$  is, assuming that the sample variance in the estimates yielded by

Algorithm 1 is  $\sigma^2$ , is in

$$\tilde{\mathcal{O}} \left( \frac{\sigma^2}{\epsilon^2} \left( \frac{|E|d_{k-1}}{|Cl_k(G)|} + \frac{D^4 \kappa^3}{\gamma_M \epsilon} \left( \log(d_{k-1})D^{-2} + \frac{\kappa^3}{\epsilon} \right) \right) \right), \quad (90)$$

where  $\kappa$  is the ratio of the largest eigenvalue to the smallest nonzero eigenvalue, i.e., the condition number of the combinatorial Laplacian,  $|E|$  is the size of the edge set in the input graph, and  $|Cl_k(G)|$  is the number of  $k$ -cliques in the input graph (this is *not* equal to  $d_{k-1}$  in general!). This shows that even in cases where the dimension is exponentially large, we can use path integration to estimate the ratio  $\dim \ker \Delta_k / d_{k-1}$  using a number of operations that scales polynomially with the number of vertices  $n$  provided that  $\sigma$ ,  $D$ ,  $\kappa$ ,  $\gamma_M^{-1}$ , and  $|Cl_k(G)|^{-1}$  are at most poly( $n$ ). We also show that  $\sigma$  can be polynomially large in some cases in Appendix G.

Quantum algorithms for TDA were thought to outperform classical counterparts in the clique dense case [15], but this algorithm serves as a counterexample. Another key point behind this dequantization result is that while an exponentially large dimension is a necessary condition for an exponential speedup for quantum TDA, it is not a sufficient condition. This implies that further work is needed in order to understand when, and even if, quantum algorithms can provide truly exponential advantages relative to all classical randomized algorithms for TDA.

Apers *et al.* [36] subsequently developed another classical path-integral Monte Carlo algorithm, which can also be efficient in some of the regimes where the quantum algorithm for Betti-number estimation works best. It is therefore natural to investigate how it fares compared to our classical and quantum algorithms. They study additive error estimation of the normalized Betti number, so we focus on comparing the runtime of their algorithm for this case.

Apers *et al.* claim a runtime of  $n^{\mathcal{O}(\gamma^{-1} \log(1/\epsilon))}$  for general simplicial complexes, where  $n$  is the number of vertices,  $\epsilon$  is the additive error, and  $\gamma \leq \lambda_2(\Delta_k) / \hat{\lambda}$  with  $\hat{\lambda} \in \Theta(\lambda_{\max}(\Delta_k))$ . For clique complexes in particular, their runtime is generically poly( $n$ )( $n/\hat{\lambda}$ ) $^{\mathcal{O}(\gamma^{-1} \log(1/\epsilon))}$ , or  $2^{\mathcal{O}(\gamma^{-1} \log(1/\epsilon))}$  when  $k$  or the maximum up-degree (see the last section of Appendix G for the definition) of the  $k$ -simplices are  $O(n)$ .

For general simplicial complexes, their runtime depends exponentially on  $\gamma^{-1}$  of  $\Delta_k$ , which is an upper bound on the condition number  $\kappa$ . They thus require constant  $\gamma$

---

**Data:**  $k > 0, n > 0, N_{\text{samp}} > 0, t \geq 0, r \geq 0$ , a function  $\text{Pr}(\Gamma)$  which assigns a non-zero probability to each vector  $\Gamma \in \mathbb{R}^{2rD}$ , a function  $W(\Gamma) = \langle \lambda_{p_1, \Gamma_1} | \lambda_{p_2, \Gamma_2} \rangle \cdots \langle \lambda_{p_{2rD-1}, \Gamma_{2rD-1}} | \lambda_{p_1, \Gamma_1} \rangle$  where  $|\lambda_{p_j, \Gamma_j}\rangle$  is the  $\Gamma_j^{\text{th}}$  eigenvector of the one sparse matrix  $U_{p_j}$ .

**Result:** Estimate  $\bar{E}$  which is an unbiased estimator of  $\beta_{k-1}/d_{k-1}$

**for**  $q$  **from** 1 **to**  $N_{\text{samp}}$  **do**

$\Sigma \leftarrow$  a set of  $k$  points encoded as an integer

**while**  $\Sigma$  is not a  $(k-1)$ -simplex **do**

$\Sigma \leftarrow$  a random set of  $k$  points encoded as an integer

**end**

Draw a vector  $\Gamma = [\Sigma, \Gamma_2, \dots, \Gamma_{2rD}]$  from the probability distribution  $\text{Pr}(\Gamma)$ .

$E_q \leftarrow \frac{1}{d_{k-1}} \left( \frac{\exp(-\lambda_{p_1, \Gamma_1} t / r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t / 2r) W}{\text{Pr}(\Gamma)} \right)$

**end**

$\bar{E} \leftarrow \frac{1}{N_{\text{samp}}} \sum_q E_q$  average of  $E$ .

---

Algorithm 1. Classical randomized algorithm for Betti-number computation.

and  $\epsilon$  for general simplicial complexes. By contrast, our classical algorithm above has a runtime depending on fixed polynomials in  $\gamma^{-1}$  and  $\epsilon^{-1}$ , and can thus tolerate poly( $n$ ) scaling for both provided  $\sigma^2$  also depends polynomially on  $\kappa$  and  $1/\epsilon$ . For clique complexes, their best-case runtime is polynomial if  $\gamma \in \Omega(1)$  and  $\epsilon = 1/\text{poly}(n)$ , or  $\gamma \in \Omega(1/\log n)$  and  $\epsilon \in \Omega(1)$ . Analogous conclusions hold for our classical algorithm in these regimes if we are once again given suitable promises on the polynomial dependence of  $\sigma^2$  on those parameters. Thus given the difficulty in analyzing the dependence of  $\sigma^2$  on  $\kappa$  and  $\epsilon^{-1}$  in general, our algorithms cannot be easily compared. In the worst-case bound for  $\sigma^2$  given in Eq. (G37), we would require  $D\kappa$  to be poly-logarithmic in  $n$  and  $\kappa/\epsilon$  a constant at worst and thus also cannot tolerate inverse polynomial error scaling or spectral gap.

The algorithm of Apers *et al.* can efficiently compute a *constant* additive error estimate of the normalized Betti number but cannot efficiently compute an *inverse polynomial* additive error estimate for the graphs  $K(n/k, k)$  discussed in this work. This yields an immediate exponential separation with the quantum algorithm since the latter is able to efficiently compute such an inverse polynomial additive error estimate. Moreover, even stronger separations are possible. As noted above, the runtime of the algorithm of Apers *et al.* depends exponentially on the inverse of the (normalized) spectral gap of the combinatorial Laplacian, unlike in the case of the quantum algorithm whose runtime depends only polynomially on this parameter.

For the family  $K(n/k, k)$ , this (normalized) spectral gap is large and suitable for the algorithm of Apers *et al.* There are nevertheless many graphs that do have a (normalized) spectral gap, which incurs an additional exponential scaling for the case of the classical algorithm but not for the quantum one. For example, by adding a single edge to each cluster in  $K(n/k, k)$  the (normalized) spectral gap of the combinatorial Laplacian becomes much smaller, which

causes the algorithm of Apers *et al.* to no longer be able to efficiently compute even a constant additive precision estimate of the normalized Betti number. On the other hand, the quantum algorithm can still efficiently compute an inverse polynomial additive error estimate as can our classical algorithm given similar guarantees on the polynomial scaling of  $\sigma^2$  with  $1/\gamma$ . In Appendix F 1, we detail some of these modified graph examples with smaller spectral gaps.

## V. CONCLUSION

In order to provide applications where quantum computers can practically outperform classical computers on hardware anticipated in the near future, it is necessary to develop algorithms where there is a greater than square-root speedup in the complexity [7]. This is because the large overheads involved in implementing quantum gates in an error-corrected code mean that there is a huge slowdown in the gate frequency as compared to classical computers. When the Betti number is of order 1, the complexity of the quantum algorithm for estimating Betti numbers is only a square-root speedup over classical approaches. This is as compared to classical approaches that scale approximately linearly in  $\binom{n}{k}$ . On the other hand, when the Betti number is large, the quantum complexity of estimating the Betti number to given *relative* error (error as a ratio to the Betti number) will be small.

There exist classes of graphs with very large Betti numbers. We introduce graph classes, which exhibit Betti numbers in the regime where the speedup is superpolynomial; specifically it is approximately a  $2 \ln n$  root. The magnitude of the speedup is limited by the need to enter the data in the quantum algorithm, which introduces a  $|E|$  factor to the complexity. These are very specially constructed graphs for large Betti number, but we show there exist far more general classes of graphs whose parameters give a *quartic* speedup over naive classical algorithms, showing that speedups beyond quadratic are possible far more generally.

We have also provided a host of new techniques for quantum Betti-number estimation that reduce the complexity. These include Kaiser-window amplitude estimation, improved Dicke state preparation, and improved eigenstate filtering. These improvements greatly improve the complexity in many ways, though the main scaling of the complexity as  $\sqrt{\binom{n}{k}}$  remains. In particular, we have major improvements in the complexity arising from improved Dicke state preparation as well as improved amplitude estimation, which together give about 3 orders of magnitude improvement. We have further improvements arising from our clique-checking procedure, separation of amplitude estimation and amplification in initial state preparation, and use of filtering instead of phase estimation, which may give a further order of magnitude of improvement depending on how prior work is interpreted. Moreover, our methods enable accurate estimation of the complexity of the quantum algorithm, including all constant factors.

Based on that, we estimate that tens of billions of Toffolis would be sufficient to estimate a Betti number that should be classically intractable. This number of Toffoli gates is reasonable for early generations of fully fault-tolerant quantum computers. While the exact threshold for quantum advantage depends on constant factors in the classical algorithms, it seems likely that this application will fall somewhere in between quantum chemistry applications and Shor’s algorithm in terms of the resources required for quantum advantage. The standard classical approaches would be expected to be intractable because they would require an extremely large storage. We have also presented an alternative approach for classical estimation that could be more efficient, because it does not require large storage and instead requires Monte Carlo sampling. That classical approach may be tractable, but it is difficult to evaluate its complexity because it depends on the gap of a Markov chain, which is unknown.

There is scope for further improvement of the quantum algorithm for Betti numbers by implementing a more efficient method of clique finding. We have currently applied just amplitude amplification for clique finding, but there are more efficient classical methods for clique finding that could potentially be adapted for the quantum algorithm. That is nontrivial because these methods often require large storage, which would not be practical in the quantum algorithm where we need to minimize the number of ancilla qubits.

## ACKNOWLEDGMENTS

The authors acknowledge helpful conversations with David Gamarnik, Robin Kothari, Seth Lloyd, Alexander Schmidhuber, Nikhil Srivastava, and Adam Zalcman. D.W.B. worked on this project under a sponsored research agreement with Google Quantum AI. D.W.B. is also supported by Australian Research Council Discovery Projects

DP190102633 and DP210101367. N.W. and V.D. were funded by a grant from Google Quantum AI. N.W. was also funded by grants from the US Department of Energy, Office of Science, National Quantum Information Science Research Centers, Co-Design Center for Quantum Advantage under Contract No. DE-SC0012704. V.D. and C.G. were supported by the Dutch Research Council (NWO/ OCW), as part of the Quantum Software Consortium programme (Project No. 024.003.037). Some of the discussions and collaboration for this project occurred while using facilities at the Kavli Institute for Theoretical Physics, supported in part by the National Science Foundation under Grant No. NSF PHY-1748958. This work was also supported by the Dutch National Growth Fund (NGF), as part of the Quantum Delta NL programme.

## APPENDIX A: DETAILED BACKGROUND ON TOPOLOGICAL DATA ANALYSIS

We present some background material from singular homology needed in topological data analysis, broadly following the treatments in Refs. [47–49].

Let  $v_0, \dots, v_k$  be  $k + 1$  distinct points in  $\mathbb{R}^n$ . The set  $\{v_0, \dots, v_k\}$  is said to be **affinely independent** if the set  $\{v_1 - v_0, \dots, v_k - v_0\}$  is linearly independent. In other words, we consider the given set of points to be affinely independent if when we take one of the points to be the “origin” (say  $v_0$  without loss of generality) and draw vectors from this point to the others, the collection of the resulting vectors is linearly independent.

If  $\{v_0, \dots, v_k\}$  is affinely independent, the **k-simplex** spanned by them is the set

$$[v_0, \dots, v_k] := \left\{ \sum_{i=0}^k t_i v_i : t_i \geq 0 \text{ and } \sum_{i=0}^k t_i = 1 \right\}. \quad (\text{A1})$$

Equivalently, a simplex is just the convex hull of its affinely independent set of vertices. The points  $v_i$  are the **vertices** of the simplex and the integer  $k$  is the **dimension** of the simplex. Figure 4 shows some examples of simplices. Note that it follows from the definitions given that  $k \leq n$  since any set of  $n + 2$  points or more cannot be affinely independent. This is because no collection of  $n + 1$  vectors or more in an  $n$ -dimensional vector space can be linearly independent. Such a collection of vectors therefore cannot determine any simplices of dimension  $n + 1$  or higher.



FIG. 4. A 0-simplex (point), 1-simplex (edge), 2-simplex (triangle), and 3-simplex (tetrahedron) from left to right.

Let  $\sigma$  be a  $k$ -simplex. A simplex spanned by a nonempty subset of the vertices of  $\sigma$  is a **face** of  $\sigma$ . For example, the 0-dimensional faces of  $\sigma$  are its vertices and its one-dimensional faces are the edges, which are spanned by two vertices. Faces of  $\sigma$  that are not equal to  $\sigma$  are called **proper faces**. The  $(k-1)$ -dimensional faces of  $\sigma$  are called its **boundary faces** and their union is its **boundary**.

**Definition 2 (Simplicial Complex).**—A simplicial complex  $S$  is a finite collection of simplices satisfying the following conditions:

- (1) If  $\sigma \in S$ , every face of  $\sigma$  is in  $S$ .
- (2) If  $\sigma_1, \sigma_2 \in S$ , then  $\sigma_1 \cap \sigma_2 = \emptyset$  or  $\sigma_1 \cap \sigma_2$  is a face of  $\sigma_1$  and  $\sigma_2$ .

The first condition says that a simplicial complex should also contain all the faces of a given simplex in the complex. The second condition says that any two simplices in a simplicial complex either do not intersect or intersect at a common face of both. Examples of a simplicial complex and a similar object which does not satisfy the conditions are given in Fig. 5. We define the **dimension of a simplicial complex** to be the maximum of the dimensions of all simplices in the complex.

It can be shown that a simplicial complex is completely determined by its vertices and information about which sets of vertices span which simplices. This provides the motivation for the following definition.

**Definition 3 (Abstract Simplicial Complex).**—An abstract simplicial complex is a collection  $C$  of nonempty finite sets such that if  $s \in C$ , then every nonempty subset of  $s$  is also in  $C$ .

This general notion of a simplicial complex is particularly useful when we wish to construct one “abstractly,” i.e., without reference to a particular embedding into Euclidean space.

We will mainly be concerned with computing topological invariants of a certain kind of simplicial complexes constructed from graphs, called clique complexes, throughout this paper. The graphs serving as inputs into the quantum algorithms in this paper are not necessarily induced by any finite-dimensional data and are instead

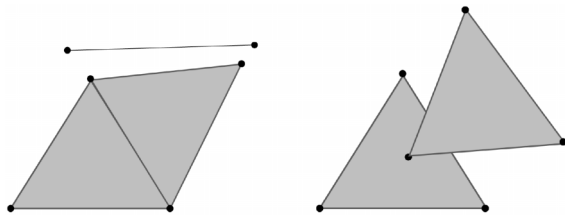


FIG. 5. (Left) A 2D simplicial complex in  $\mathbb{R}^2$ . (Right) A set that is not a simplicial complex in  $\mathbb{R}^2$ . It violates condition 2 of Definition 2.

abstract graphs, i.e., abstract simplicial complexes in the sense of Definition 3 in the preceding section.

**Definition 4 (Graphs).**—A graph  $G$  is a pair of objects  $G = (V, E)$ , where  $V$  is a set of elements referred to as the “vertices” of  $G$  and  $E$  is a set consisting of pairs of vertices thought of as “edges” connecting the pairs of vertices.

Given an undirected graph (i.e., in which the edges are not assumed to have direction), a **clique**  $C$  of a graph is a subset of  $V$  such that every pair of distinct vertices in  $C$  is connected by an edge.  $C$  is called a  $k$ -clique if  $|C| = k$ .

We now define the notion of a clique complex.

**Definition 5 (Clique Complex).**—The clique complex of a graph  $G$  is the abstract simplicial complex formed by associating a  $k$ -simplex to every  $k+1$ -clique in  $G$ .

The invariants we are most interested in are the Betti numbers of a clique complex associated to a graph, which give the number of holes of a given dimension in that clique complex. We now show how to determine the Betti numbers of an arbitrary simplicial complex via simplicial homology.

Let  $K$  be a simplicial complex and consider the set of  $k$ -simplices in  $K$ . In what follows, we would like to make sense of taking “linear combinations” of the  $k$ -simplices in  $K$  with coefficients in some field  $R$  (we will need only  $R = \mathbb{R}$  or  $R = \mathbb{C}$  for our purposes). A  **$k$ -chain** is a formal sum of  $k$ -simplices  $\sum_i c_i \sigma_i$  where  $\sigma_i \in K$ ,  $c_i \in R$ . The set of all  $k$ -chains is denoted by  $C_k(K)$  and is a vector space over  $R$ . The  $k$ -simplices form a basis for  $C_k(K)$ , so the dimension of  $C_k(K)$  equals the number of  $k$ -simplices in  $K$ .

**Definition 6 (Boundary Map).**—Let  $\sigma = [v_0, \dots, v_k]$  be a  $k$ -simplex. The boundary map on  $k$ -simplices is a map

$$\partial_k: C_k(X) \rightarrow C_{k-1}(X) \quad (\text{A2})$$

that acts as

$$\partial_k \sigma = \sum_{i=0}^k (-1)^i [u_0, u_1, \dots, \hat{u}_i, \dots, u_k], \quad (\text{A3})$$

where  $\hat{u}_i$  denotes that the vertex  $i$  has been removed.

The boundary map acts on  $k$ -simplices  $\sigma \in C_k(K)$  and gives a  $(k-1)$ -simplex  $\partial_k \sigma$  that can be interpreted as the boundary of  $\sigma$ .

An  **$k$ -cycle** is a  $k$ -chain  $c \in C_k(K)$  such that  $\partial_k c = 0$ . Therefore,  $k$ -cycles are precisely the kernel of the boundary map and are a subspace of  $C_k(K)$  denoted by  $Z_k = \ker \partial_k$ . An  $k$ -chain  $c$  is an  **$k$ -boundary** if there exists an  $(k+1)$ -chain  $\sigma \in C_{k+1}(K)$  such that  $c = \partial_{k+1}(\sigma)$ . Equivalently,  $k$ -boundaries are precisely the image of the boundary map and form a subspace denoted by  $B_k(K) = \text{Im } \partial_{k+1}$ . Figure 6 shows an example of a 1-boundary and 1-cycle.

It turns out that there is a relationship between the two subspaces as implied by the following fundamental result in homology theory.



FIG. 6. (Left) The 1-boundary of a 2-chain as indicated by the arrows. (Right) A 1-cycle that is not the boundary of any 2-chain.

**Proposition 5.**— $\partial_k \circ \partial_{k+1}(\sigma) = 0$  for all  $k + 1$  chains  $\sigma$  and all  $0 \leq k \leq \dim K$ .

The above proposition essentially says that the boundary of a boundary is 0. It implies that the image of  $\partial_{k+1}$  is contained in the kernel of  $\partial_k$ . This allows us to define the homology groups as follows:

**Definition 7 (Homology Groups).**—The  $k$ th singular homology group  $H_k$  of a simplicial complex is the quotient vector space

$$H_k(K) = Z_k(K)/B_k(K) = \text{Ker } \partial_k / \text{Im } \partial_{k+1}. \quad (\text{A4})$$

**Definition 8 (Betti Numbers).**—The  $i$ th Betti number  $\beta_i$  is the dimension of the  $i$ th homology group.

The  $i$ th homology group is generated by cycles that are not the boundaries of any simplex. In other words, these are simplices that enclose a “void” or “hole” (see the right image of Fig. 6 above). It is worth noting that  $\beta_0$ , the 0th Betti number, represents the number of connected components the simplicial complex has.

The problem of computing the Betti numbers of a simplicial complex therefore reduces to the problem of computing the rank of the boundary map. A common and simple classical approach to doing this at the computational level is as follows.

Let  $K$  be a simplicial complex and assume for simplicity we are working over the field  $\mathbb{Z}_2$ . Label the  $p$ -simplices in  $C_p(K)$  by  $x_1, \dots, x_{n_p}$  and the  $(p - 1)$ -simplices in  $C_{p-1}(K)$  by  $y_1, \dots, y_{n_{p-1}}$ . These simplices form bases for  $C_p(K)$  and  $C_{p-1}(K)$  as mentioned previously. We can then represent the action of the boundary map  $\partial_p$  on  $C_p(K)$  as follows:

$$\partial_p(x_j) = \sum_{i=1}^{n_{p-1}} a_j^i y_i \text{ where } a_j^i = \begin{cases} 1 & \text{if } y_i \text{ is a face of } x_j \\ 0 & \text{otherwise} \end{cases}. \quad (\text{A5})$$

Then for any  $p$ -chain  $c = \sum_{j=1}^{n_p} a_j x_j$ , we can write the above in matrix form

$$\partial_p c = \begin{bmatrix} a_1^1 & a_1^2 & \dots & a_1^{n_p} \\ a_2^1 & a_2^2 & \dots & a_2^{n_p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_{p-1}}^1 & a_{n_{p-1}}^2 & \dots & a_{n_{p-1}}^{n_p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n_p} \end{bmatrix}. \quad (\text{A6})$$

Thus the boundary map on  $C_p(K)$  can be represented as an  $n_{p-1} \times n_p$  sparse matrix with entries in  $\mathbb{Z}_2$ . The columns of this matrix span  $\text{Im } \partial_p = B_{p-1}$ , so  $\text{rank } \partial_p = \dim B_{p-1} = b_{p-1}$ . The boundary matrix can be brought into the Smith normal form via a generalization of Gaussian elimination, which applies to any principal ideal domain (this includes fields). This results in a matrix with a number of 1’s on the diagonal equal to the rank of the boundary matrix. Gaussian elimination for  $\partial_p$  takes  $\mathcal{O}(n_{p-1} n_p \min(n_p, n_{p-1}))$  time and requires  $\mathcal{O}((n_{p-1} + n_p)^2)$  memory [50,51]. When working over arbitrary fields, the same procedure holds with the exception that the matrix elements of the boundary operator can be  $\pm 1$  or 0.

In the worst-case scenario, however, the number of  $p$ -simplices grows exponentially with the number of points in the complex, so this procedure becomes intractable for large data sets. This motivates the problem of finding efficient classical and quantum algorithms for extracting Betti numbers of simplicial complexes.

## APPENDIX B: SCHEMES FOR DICKE STATE PREPARATION

### 1. First scheme

In our first scheme we prepare  $n$  registers with  $n_{\text{seed}} := \lceil \log cn \rceil$  qubits in equal superposition. Then the principle is to find a threshold such that  $k$  registers are less than or equal to this threshold, and the inequality test is used to set the ones in the Dicke state. In the limit of large  $n_{\text{seed}}$ , there is very low amplitude for the registers to be equal to each other, which can cause a failure. The problem is that making  $n_{\text{seed}}$  larger increases the complexity. Reducing  $n_{\text{seed}}$  will reduce the complexity but increase the amplitude for failure. We will adjust the value of  $c$  in order to make both the complexity and probability of failure reasonably small.

The steps of the procedure are as follows.

- (1) Prepare  $n$  registers in an equal superposition of  $2^{n_{\text{seed}}}$  values. This may be done with Hadamards (no non-Clifford gates).
- (2) Sum the most significant bits of these registers. This may be done with no more than  $n$  Toffolis. This gives the number of registers at least as large as  $10\,000 \dots$
- (3) Perform an inequality test with  $k$  (with  $\log n$  Toffolis) and place the result in an ancilla. That is, we are testing if the number is  $\geq k$ , and we call the bit representing the result  $b_1$ .
- (4) For  $j = 2$  to  $n_{\text{seed}}$ , we perform the following.
  - (a) Perform an inequality test on each register checking if the first  $j$  bits are  $\geq b_1 \dots b_{j-1}$ . The cost is  $(j - 1)n$  Toffolis.
  - (b) Sum the results of those inequality tests, with cost  $n$ .



- (c) Perform an inequality test if that number is  $\geq k$ , setting the result as  $b_j$ , with cost  $\log n$ .
- (5) Perform inequality tests on each register to check if all  $n_{\text{seed}}$  bits are  $\geq b_1 \dots b_{n_{\text{seed}}}$ . The cost is  $n_{\text{seed}}n$  Toffolis.
- (6) Sum the results of those inequality tests, with cost  $n$ .
- (7) Check if the result is equal to  $k$ , with cost  $\log n$ .

The logic of this procedure is that whenever there are more than  $k$  registers greater than or equal to the threshold, we increase the threshold, which is given by the bits  $b_1 b_2 \dots$ , otherwise we reduce it. This test corresponds to the inequality test  $\geq k$  in part (4)(c), and the threshold is adjusted by choosing  $b_j$ . The ones for the Dicke state are given by the inequality tests in step (5). At the end, provided we have success where the sum of the ones checked in step (7) was equal to  $k$ , then we have  $k$  ones in a superposition of permutations corresponding to a Dicke state. The prepared state is entangled with the ancilla registers, but this is suitable for our application. Summing all the costs in this procedure gives a total Toffoli cost

$$(n_{\text{seed}} + 1) \left[ \frac{n}{2} (n_{\text{seed}} + 2) + \lceil \log n \rceil \right]. \quad (\text{B1})$$

To derive the probability of success, note that the failure occurs where it is not possible to provide a threshold where there are  $k$  numbers greater than or equal to it. That will be true if the  $k$ th smallest and  $k + 1$ th smallest numbers are equal. In other words, if we were to sort our numbers, and compare the values in register  $k$  and register  $k + 1$ , there would be failure if these numbers were equal. (Note that we do not perform this sort in practice.)

The success probability for a given  $k$  may be given in the form of a sum as

$$\frac{1}{(cn)^n} \binom{n}{k} \sum_{\ell=1}^{cn} [\ell^k - (\ell - 1)^k] (cn - \ell)^{n-k}, \quad (\text{B2})$$

where for simplicity we are choosing  $c$  such that  $cn$  is a power of 2. This expression is obtained by considering values up to  $\ell$  in registers 1 to  $k$ , and greater than  $\ell$  for registers  $k + 1$  to  $n$ . The number of combinations of  $k$  registers with a maximum of  $\ell$  is  $\ell^k - (\ell - 1)^k$ , and the number of combinations of  $n - k$  registers with values above  $\ell$  is  $(cn - \ell)^{n-k}$ . The factor of  $\binom{n}{k}$  is to account for the number of positions to choose the  $k$  registers with maximum value  $\ell$ . Summing over  $\ell$  then gives the number of combinations of values where there is a success when sorting the numbers, and we divide by the total number  $(cn)^n$  to give the probability.

This sum can be approximated by an integral to give

$$\begin{aligned} & \binom{n}{k} \sum_{\ell=0}^{cn} \left( \frac{\ell}{cn} \right)^k \left( 1 - \frac{\ell}{cn} \right)^{n-k} \\ & - \binom{n}{k} \sum_{\ell=1}^{cn} \left( \frac{\ell}{cn} - \frac{1}{cn} \right)^k \left( 1 - \frac{\ell}{cn} \right)^{n-k} \\ & \approx cn \binom{n}{k} \int_0^1 dx x^k (1-x)^{n-k} - cn \binom{n}{k} \\ & \quad \times \int_{1/cn}^1 dx \left( x - \frac{1}{cn} \right)^k (1-x)^{n-k} \\ & = \frac{cn}{n+1} - \frac{(cn-1)^{n+1}}{(cn)^n (n+1)} \\ & = \frac{cn}{n+1} \left[ 1 - \left( 1 - \frac{1}{cn} \right)^{n+1} \right] \\ & \geq 1 - \frac{1}{2c}. \end{aligned} \quad (\text{B3})$$

The integrals are evaluated by repeated integration by parts, which cancels the factor of  $\binom{n}{k}$ . This shows that the exact expression is well approximated by an upper bound of  $1 - 1/2c$ . See Appendix C for analysis of the error in this approximation, showing that it is of order  $\mathcal{O}(c^{-(\min(k, n-k)+1)})$ . Moreover, we show that  $1 - 1/2c$  is an exact lower bound to the success probability.

In practice, we find that moderate constant values of  $c$  are suitable for minimizing the complexity, and we will take  $c = 8$  in the examples below. That only increases the cost about 3%, while only needing another three qubits to store the numbers. The result of taking  $c$  constant is that the Toffoli complexity is approximately  $(n/2) \log^2 n$ . In comparison, in the sorting approach from Ref. [24], the complexity is approximately  $2n_{\text{seed}}n \log n$ , where  $n \log n$  is the number of steps in the sorting network. Moreover, that approach fails when *any* of the numbers in the sorted registers have equal values, which requires taking  $2^{n_{\text{seed}}} \gtrsim n^2$  to provide a reasonable probability of success. That results in a preparation complexity of approximately  $4n \log^2 n$ , or 8 times what we have provided here.

As a simple example where our threshold procedure is performed, consider  $n = 4$  and  $c = 4$ , so we have four registers of four qubits. An example of a basis state is

$$|0110\rangle |1110\rangle |0111\rangle |0010\rangle. \quad (\text{B4})$$

Now say we aim for  $k = 2$ . The steps are as follows.

- (a) Sum the first (most significant bits) giving 1. This is equivalent to checking how many registers are at least 1000. The sum is 1, which is less than  $k$ , so  $b_1 = 0$ .

- (b) Perform an inequality test on the first two bits of each register with 01. We get a total of 3, which is greater than  $k$ , so  $b_2 = 1$ .
- (c) Perform an inequality test of the first three bits with 011. We get 3 again, so  $b_3 = 1$ .
- (d) Perform an inequality test with 0111. We get a total of 2, which is equal to  $k$ , so we get  $b_4 = 1$ .
- (e) Perform an inequality test with 0111 again. Now test equality with  $k$ , which succeeds, so we get overall success.

As an alternative to the above example, consider the case that the basis state is

$$|0110\rangle |1110\rangle |0110\rangle |0010\rangle. \quad (\text{B5})$$

Then in the second-last step, the total would be 1, so we would have  $b_4 = 0$ , and in the last step we would test inequality with 0110. That would give three greater than or equal to 0110, so we would fail the test that the number is equal to  $k$ . In this case, if you sort the numbers you have 0010, 0110, 0110, 1110, and you can see that the second and third numbers are equal. This is the condition for failure discussed above.

## 2. Second scheme

Here we provide an alternative scheme for Dicke state preparation that can be more efficient, but provides a lower amplitude for success. It is useful in the case of large  $n$  but small  $k$  (smaller than  $\sqrt{n}$ ), where  $\binom{n}{k} \sim n^k/k!$ . The steps of this scheme are as follows.

- (1) First prepare  $k$  registers in equal superposition states over  $n$  values. The complexity of preparing an equal superposition over  $n$  values (with high probability of success) is  $4\lceil \log n \rceil + 1$  Toffolis [6], so this would give complexity  $k(4\lceil \log n \rceil + 1)$ . In the case where  $n$  is a power of 2, then this preparation can be performed just with Hadamards.
- (2) For each of the  $k$  registers, apply unary iteration as in Ref. [30] with the  $n$  qubits for the Dicke state as the target. This is used to flip the corresponding qubit. The complexity of each unary iteration is  $n - 2$ , giving total complexity  $k(n - 2)$ . There will be approximately  $\log n$  temporary ancillas used in the unary iteration that are reset to zero.
- (3) Now we sum the bits in the string of  $n$  qubits. A method of summing bits is given in Ref. [6], where multiple groups of bits are summed, and their sums are summed. The overall complexity is no more than  $2n$  Toffolis, and only a logarithmic number of ancillas is used. The ancillas used that need to be kept for later stages of the calculation can be given as  $\lceil \log n \rceil$ , and there are  $2\lceil \log n \rceil$  temporary ancillas used.

- (4) The sum of the bits is compared to  $k$ . This has complexity  $\lceil \log k \rceil$  because we are guaranteed that the number of ones is at most  $k$ .

As before, this is a scheme which prepares the Dicke state in an entangled state with an ancilla. The overall complexity is then

$$(k + 2)n + k(4\lceil \log n \rceil - 1) + \lceil \log k \rceil \quad (\text{B6})$$

Toffolis with the number of ancillas used being

$$(k + 1)\lceil \log n \rceil \quad (\text{B7})$$

with a logarithmic number of working ancillas as well. In comparison, the scheme used above uses a number of qubits scaling as  $n \log n$ , so is much larger when  $n \gg k$ . In the case where  $n$  is a power of 2, the initial preparation of the equal superposition can be just performed with Hadamards, and so the Toffoli complexity is

$$(k + 2)n - 2k + \lceil \log k \rceil. \quad (\text{B8})$$

There is a probability of failure for the preparation, because it is possible for the ones to overlap, which will be detected in the final stage where the sum of bits is compared to  $k$ . This is an example of the birthday problem, and the probability of success will be at least 1/2 when  $k$  is not larger than approximately  $\sqrt{n}$ . The probability of success is more specifically given by

$$\frac{k!}{n^k} \binom{n}{k}. \quad (\text{B9})$$

It is possible to perform amplitude amplification on this step, but it is simpler to combine this step with the clique finding. The net result on the complexity is that wherever we have  $\binom{n}{k}$  in the original costs, it is replaced with  $n^k/k!$ .

## APPENDIX C: ACCURACY OF DICKE PREPARATION SUCCESS PROBABILITY

Here we analyze the accuracy of the approximation of the sums by integrals in Eq. (B3). The sums correspond to the integrals discretized in steps of  $1/cn$ , so the approximation can be expected to converge for large  $c$ . Even for moderate  $c$  the approximation of the sums by integrals is highly accurate, because the integrands are zero at the bounds of the integrals, and the first nonzero derivatives at the bounds are of order  $\min(k, n - k)$ . That means the error term in the Euler-Maclaurin formula will correspond to the derivative of order  $\min(k, n - k)$ .

In particular, since the summands have nonzero derivatives only up to order  $n$ , the first sum can be given by

$$\sum_{\ell=0}^{cn} \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} = \int_0^{cn} d\ell \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} + \frac{f(0) + f(cn)}{2} + \sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} \frac{B_{2j}}{(2j)!} [f^{2j-1}(cn) - f^{2j-1}(0)], \quad (C1)$$

where  $f(\ell)$  is the summand,  $B_{2j}$  are Bernoulli numbers, and the upper bound on the sum over  $j$  is chosen so that only derivatives up to order  $n$  are included. Now the summand is zero at the bounds,  $f^{2j-1}(0)$  is nonzero only for  $2j - 1 \geq k$ , and  $f^{2j-1}(cn)$  is nonzero only for  $2j - 1 \geq n - k$ . When  $2j - 1 \geq k$ , the only nonzero term in the derivative at  $\ell = 0$  is for  $k$  derivatives of  $(\ell/cn)^k$  and  $2j - 1 - k$  derivatives of  $(1 - \ell/cn)^{n-k}$ . This is because we need exactly  $k$  derivatives of  $(\ell/cn)^k$  for it to be nonzero at  $\ell = 0$ . The general Leibniz rule therefore gives

$$\begin{aligned} \left. \frac{d^{2j-1}}{d\ell^{2j-1}} \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} \right|_{\ell=0} &= \binom{2j-1}{k} \left\{ \frac{d^k}{d\ell^k} \left(\frac{\ell}{cn}\right)^k \right\} \left\{ \frac{d^{2j-1-k}}{d\ell^{2j-1-k}} \left(1 - \frac{\ell}{cn}\right)^{n-k} \right\} \Big|_{\ell=0} \\ &= (-1)^{2j-1-k} \frac{(2j-1)!}{(cn)^{2j-1} k! (2j-1-k)!} k! \frac{(n-k)!}{[n-k-(2j-1-k)]!} \\ &= -(-1)^k \frac{(2j-1)!(n-k)!}{(cn)^{2j-1} (2j-1-k)!(n-2j+1)!}. \end{aligned} \quad (C2)$$

Similarly, the only nonzero term in the derivative at  $\ell = cn$  is for  $n - k$  derivatives of  $(1 - \ell/cn)^{n-k}$ , and  $2j - 1 - n + k$  derivatives of  $(\ell/cn)^k$ . In that case the general Leibniz rule gives

$$\begin{aligned} \left. \frac{d^{2j-1}}{d\ell^{2j-1}} \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} \right|_{\ell=cn} &= \binom{2j-1}{n-k} \left\{ \frac{d^{2j-1-n+k}}{d\ell^{2j-1-n+k}} \left(\frac{\ell}{cn}\right)^k \right\} \left\{ \frac{d^{n-k}}{d\ell^{n-k}} \left(1 - \frac{\ell}{cn}\right)^{n-k} \right\} \Big|_{\ell=cn} \\ &= (-1)^{n-k} \frac{(2j-1)!}{(cn)^{2j-1} (n-k)!(2j-1-n+k)!} \frac{k!}{[k-(2j-1-n+k)]!} (n-k)! \\ &= (-1)^{n-k} \frac{(2j-1)!k!}{(cn)^{2j-1} (2j-1-n+k)!(n-2j+1)!}. \end{aligned} \quad (C3)$$

These values for the derivatives then give us

$$\begin{aligned} \sum_{\ell=0}^{cn} \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} &= \int_0^{cn} d\ell \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} + \sum_{j=\lfloor \frac{k+1}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} \frac{B_{2j}}{(2j)!} (-1)^k \frac{(2j-1)!(n-k)!}{(cn)^{2j-1} (2j-1-k)!(n-2j+1)!} \\ &\quad + \sum_{j=\lfloor \frac{n-k+1}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} \frac{B_{2j}}{(2j)!} (-1)^{n-k} \frac{(2j-1)!k!}{(cn)^{2j-1} (2j-1-n+k)!(n-2j+1)!} \\ &= \int_0^{cn} d\ell \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} + \sum_{j=\lfloor \frac{k+1}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} \frac{B_{2j}}{(2j)!} (-1)^k \frac{(n-k)!}{(cn)^{2j-1} (2j-1-k)!(n-2j+1)!} \\ &\quad + \sum_{j=\lfloor \frac{n-k+1}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} \frac{B_{2j}}{(2j)!} (-1)^{n-k} \frac{k!}{(cn)^{2j-1} (2j-1-n+k)!(n-2j+1)!}. \end{aligned} \quad (C4)$$

Next, for the evaluation of the second integral, we have exactly the same reasoning, except the values of the derivatives at the bounds are slightly different. For the value at  $\ell = 1$  (instead of  $\ell = 0$ ), we now have a factor of

$$\left(1 - \frac{\ell}{cn}\right)^{n-k-(2j-1-k)} \Big|_{\ell=1} = \left(1 - \frac{1}{cn}\right)^{n-2j+1}. \quad (\text{C5})$$

Similarly, for the values of the derivatives at  $\ell = cn$ , we obtain a factor of

$$\left(\frac{\ell}{cn} - \frac{1}{cn}\right)^{k-(2j-1-n+k)} \Big|_{\ell=cn} = \left(1 - \frac{1}{cn}\right)^{n-2j+1}, \quad (\text{C6})$$

which is the same.

Using these factors and combining the expressions for the two sums then gives

$$\begin{aligned} & \binom{n}{k} \sum_{\ell=0}^{cn} \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} - \binom{n}{k} \sum_{\ell=1}^{cn} \left(\frac{\ell}{cn} - \frac{1}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} \\ &= \frac{cn}{n+1} \left[1 - \left(1 - \frac{1}{cn}\right)^{n+1}\right] + \sum_{j=\lceil \frac{k+1}{2} \rceil}^{\lceil \frac{n}{2} \rceil} D(n, k, c, j) + \sum_{j=\lceil \frac{n-k+1}{2} \rceil}^{\lceil \frac{n}{2} \rceil} D(n, n-k, c, j), \end{aligned} \quad (\text{C7})$$

where

$$D(n, k, c, j) := (-1)^k \frac{n!}{k!} \frac{B_{2j}}{2j (cn)^{2j-1} (2j-1-k)! (n-2j+1)!} \left[1 - \left(1 - \frac{1}{cn}\right)^{n-2j+1}\right]. \quad (\text{C8})$$

We will show that  $|D(n, k, c, j)|$  decreases with  $j$ , so the dominant terms come from the smallest  $j$ . Moreover, we will show that the largest values come from taking  $j = 1$  with  $k = 1$  or  $n - k = 1$ , in which case it is of order  $1/c^2$ . That means the largest value the correction can take is of higher order than the failure probability.

To show that  $|D(n, k, c, j)|$  decreases with  $j$ , first note that

$$B_{2j} = \frac{(-1)^{j+1} 2(2j)!}{(2\pi)^{2j}} \zeta(2j), \quad (\text{C9})$$

so the monotonic decreasing property of the Riemann zeta function  $\zeta(2j)$  gives

$$\left| \frac{B_{2j+2}}{B_{2j}} \right| < \frac{2(j+1)(2j)}{(2\pi)^2}. \quad (\text{C10})$$

We can therefore upper bound the ratio of  $|D(n, k, c, j+1)|$  to  $|D(n, k, c, j)|$  as

$$\frac{|D(n, k, c, j+1)|}{|D(n, k, c, j)|} < \frac{(2j)(2j+1)(n-2j)(n-2j+1)}{(2\pi cn)^2 (2j-k)(2j+1-k)} < \frac{1}{(2\pi c)^2}. \quad (\text{C11})$$

Here we have used the fact that the factor in square brackets for  $D(n, k, c, j)$  is monotonically decreasing in  $j$ . Note that in the sums we use both  $D(n, k, c, j)$  and  $D(n, n-k, c, j)$ , and the same reasoning holds for both. This shows that the summands  $D(n, k, c, j)$  and  $D(n, n-k, c, j)$  rapidly decrease with  $j$ , so the value for the smallest  $j$  is dominant.

Now let us consider the size of  $D(n, k, c, j)$  for the smallest  $j$  in the sum. For  $k = 1$ , the sum starts from  $j = 1$ , which gives

$$\frac{n!}{k!} \frac{B_{2j}}{(cn)^{2j-1} (2j-1-k)! (n-2j+1)!} \left[ 1 - \left( 1 - \frac{1}{cn} \right)^{n-2j+1} \right] = \frac{1}{12c} \left[ 1 - \left( 1 - \frac{1}{cn} \right)^{n-1} \right] \leq \frac{n-1}{12nc^2}, \quad (\text{C12})$$

where we have used  $B_2 = 1/6$ . Then for  $k = 2$ , the sum starts from  $j = 2$ , which gives

$$|D(n, 2, c, 2)| = \frac{(n-1)(n-2)}{240n^2c^3} \left[ 1 - \left( 1 - \frac{1}{cn} \right)^{n-3} \right] \leq \frac{(n-1)(n-2)(n-3)}{240n^3c^4}, \quad (\text{C13})$$

where we have used  $B_4 = -1/30$ .

Thus we find that the first term in the sum is  $\mathcal{O}(c^{-(k+1)})$  for  $k = 1$  or  $2$ . Next we will consider the value as we increase  $k$  in steps of 2. Let us put  $j_k := \lceil (k+1)/2 \rceil$  for the first  $j$  in the sum. Regardless of whether  $k$  is even or odd we find  $(2j_k - 1 - k)! = 1$ . For the case of odd  $k$  we have  $2j_k = k + 1$ , so

$$\frac{|D(n, k+2, c, j_{k+2})|}{|D(n, k, c, j_k)|} < \frac{(2j_k)(2j_k+1)(n-2j)(n-2j+1)}{(2\pi cn)^2(k+1)(k+2)} < \frac{1}{(2\pi c)^2}. \quad (\text{C14})$$

This shows that the sum is decreasing when we consider steps of 2 in  $k$ . For  $k$  even we have  $2j_k = k + 2$ , so

$$\frac{|D(n, k+2, c, j_{k+2})|}{|D(n, k, c, j_k)|} < \frac{k+3}{(2\pi c)^2(k+1)} \leq \frac{5}{3(2\pi c)^2}, \quad (\text{C15})$$

which shows the sum is still decreasing.

In either case, increasing the value of  $k$  by 2 corresponds to reducing the value by at least  $c^2$ , so we find that this first term in the sum is  $\mathcal{O}(c^{-(k+1)})$  in general. Moreover, because we have shown that successive terms in the sum are smaller by factors of  $1/(2\pi c)^2$ , the entire sum is  $\mathcal{O}(c^{-(k+2)})$ . We have two sums, one with  $k$  and one with  $n-k$ , so the total of the two sums over  $j$  can be given as

$$\mathcal{O}\left(\frac{1}{c^{\min(k, n-k)+1}}\right). \quad (\text{C16})$$

These results can be used to show a lower bound of  $1 - 1/2c$  for the success probability. We will take  $n \geq 3$ , since in the trivial case  $n = 2$  we find the probability is exactly  $1 - 1/2c$ . First, we use

$$\begin{aligned} 0 &\leq (n-2)n - (n-2)(n-3)/3 - n^2/3 \\ n^2/3 &\leq (n-2)n - (n-2)(n-3)/3 \\ \frac{1}{3} &\leq \frac{(n-2)}{n} - \frac{(n-2)(n-3)}{3n^2} \\ \frac{1}{2\pi^2} &< \frac{(n-2)}{n} - \frac{(n-2)(n-3)}{3n^2} \\ \frac{1}{2\pi^2} &< \frac{c(n-2)}{n} - \frac{(n-2)(n-3)}{3n^2} \\ \frac{n-1}{nc} \frac{1}{(2\pi c)^2} &< \frac{(n-1)(n-2)}{2(nc)^2} - \frac{(n-1)(n-2)(n-3)}{6(nc)^3} \\ \frac{n-1}{nc} \left( 1 - \frac{1}{(2\pi c)^2} \right) &> \frac{n-1}{nc} - \frac{(n-1)(n-2)}{2(nc)^2} + \frac{(n-1)(n-2)(n-3)}{6(nc)^3} \\ \frac{n-1}{nc} \left( 1 - \frac{1}{(2\pi c)^2} \right) &> \left[ 1 - \left( 1 - \frac{1}{cn} \right)^{n-1} \right] \\ \frac{n-1}{6nc^2} &> \frac{1}{6c} \left[ 1 - \left( 1 - \frac{1}{cn} \right)^{n-1} \right] \frac{1}{1 - \frac{1}{(2\pi c)^2}}. \end{aligned} \quad (\text{C17})$$

In the fifth line we have used  $c \geq 1$ . Next, we lower bound the probability of success by upper bounding the sums over  $j$  by the cases with  $k = 1$  or  $n - k = 1$ , and using a factor of  $(1 - 1/(2\pi c)^2)^{-1}$  to account for the sum where each successive term is less than the previous by at least a factor of  $1/(2\pi c)^2$ . We find that

$$\begin{aligned}
& \binom{n}{k} \sum_{\ell=0}^{cn} \left(\frac{\ell}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} - \binom{n}{k} \sum_{\ell=1}^{cn} \left(\frac{\ell}{cn} - \frac{1}{cn}\right)^k \left(1 - \frac{\ell}{cn}\right)^{n-k} \\
& \geq \frac{cn}{n+1} \left[1 - \left(1 - \frac{1}{cn}\right)^{n+1}\right] - 2 \frac{1}{12c} \left[1 - \left(1 - \frac{1}{cn}\right)^{n-1}\right] \frac{1}{1 - \frac{1}{(2\pi c)^2}} \\
& \geq 1 - \frac{1}{2c} + \frac{n-1}{6nc^2} - \frac{1}{6c} \left[1 - \left(1 - \frac{1}{cn}\right)^{n-1}\right] \frac{1}{1 - \frac{1}{(2\pi c)^2}} > 1 - \frac{1}{2c}.
\end{aligned} \tag{C18}$$

In the last line we have used the bound we derived in Eq. (C17). Including the case  $n - 2$  as well, this shows that the probability of success is lower bounded by  $1 - 1/2c$  in general.

For the example where  $n = 256$  and  $c = 8$ , the integral approximation gives estimated success probability 0.94001551223575. We show the error in the integral approximation in Fig. 7 for this example. As can be expected from our analysis of the Euler-Maclaurin formula, it is found that the integral approximation is more accurate as  $\min(k, n - k)$  is increased. For  $k$  near  $n/2$  the approximation is accurate to well over 200 decimal places. For  $k = 16$ , the integral approximation is accurate to about 30 decimal places. For  $k = 2$  the error in the approximation is still less than  $10^{-6}$ , though it increases to 0.0012 for  $k = 1$  or  $n - 1$ .

We can also more simply prove that  $1 - 1/2c$  lower bounds the probability of success when averaging over  $k$ . Summing Eq. (B2) over  $k = 1$  to  $n - 1$  gives

$$\begin{aligned}
\frac{1}{(cn)^n} \sum_{k=1}^{n-1} \binom{n}{k} \sum_{\ell=1}^{cn} [\ell^k - (\ell-1)^k] (cn - \ell)^{n-k} &= \frac{1}{(cn)^n} \sum_{k=0}^n \binom{n}{k} \sum_{\ell=1}^{cn} [\ell^k - (\ell-1)^k] (cn - \ell)^{n-k} - \frac{1}{(cn)^n} \sum_{\ell=1}^{cn} [\ell^n - (\ell-1)^n] \\
&= \frac{1}{(cn)^n} \sum_{\ell=1}^{cn} [(cn)^n - (cn-1)^n] - \frac{1}{(cn)^n} (cn)^n \\
&= cn - 1 - \frac{(cn-1)^n}{(cn)^{n-1}} \geq cn - 1 - \frac{(cn)^n - n(cn)^{n-1} + n(n-1)(cn)^{n-2}}{(cn)^{n-1}} \\
&= (n-1) \left(1 - \frac{1}{2c}\right).
\end{aligned} \tag{C19}$$

Dividing by  $n - 1$  for the number of values of  $k$  then gives a lower bound of  $1 - 1/2c$  for the average success probability.

#### APPENDIX D: PROOF OF COMPLEXITY OF AMPLITUDE ESTIMATION

For the complexity of amplitude estimation, the standard approach is to use phase estimation on the Grover iterate of amplitude amplification. If there is an initial amplitude of  $a$ , then the phase of each step of amplitude amplification is  $2 \arcsin a$ . The original proposal was to use control registers in the phase estimation in an equal superposition, but of course for phase estimation that is a poor choice. Here we would like a small probability of error beyond a given confidence interval, and for that case it is better to use a prolate spheroidal window [52], which the Kaiser window provides a good approximation of.

When applying phase measurement, we would start with a control state of the form (omitting normalization)

$$\sum_{m=-N}^N \frac{1}{2N} \frac{I_0\left(\pi\alpha\sqrt{1 - (m/N)^2}\right)}{I_0(\pi\alpha)} |m\rangle, \tag{D1}$$

where  $I_0$  is a zeroth-order modified Bessel function of the first kind. If we call the operator combining  $U$  and the reflection on the flag qubit  $W$ , then we would then control between applications of  $W$  and  $W^\dagger$  with eigenvalue  $e^{i\theta}$  to give

$$\sum_{m=-N}^N \frac{1}{2N} \frac{I_0\left(\pi\alpha\sqrt{1 - (m/N)^2}\right)}{I_0(\pi\alpha)} e^{im\theta} |m\rangle. \tag{D2}$$

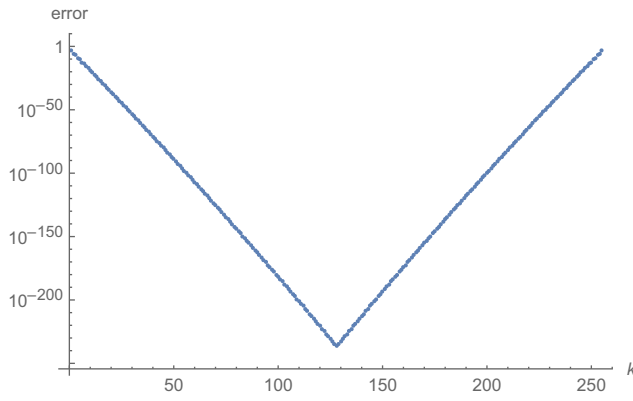


FIG. 7. The error in the integral approximation in Eq. (B3) as a function of  $k$  for  $n = 256$ ,  $c = 8$ .

The inverse quantum Fourier transform then corresponds to an inner product with a phase state

$$\frac{1}{\sqrt{2N+1}} \sum_{m=-N}^N e^{im\hat{\theta}} |m\rangle. \quad (\text{D3})$$

The inner product then gives the Fourier transform, so is proportional to

$$\frac{\sin\left(\sqrt{N^2(\hat{\theta} - \theta)^2 - (\pi\alpha)^2}\right)}{I_0(\pi\alpha)\sqrt{N^2(\hat{\theta} - \theta)^2 - (\pi\alpha)^2}}. \quad (\text{D4})$$

This needs to be squared to give the probability distribution for the error in the phase measurement.

The distribution has its first zero for  $\theta = (\pi/N)\sqrt{1+\alpha^2}$ , so to estimate the probability in the wings of the distribution we should integrate past that point. We also have the difficulty that we are not given the exact normalization of the probability distribution. To approximate the normalization, we can approximate the center of the distribution by a Gaussian. The approximation can be found by taking the Taylor series of the log of the distribution about zero, and gives

$$\frac{\sinh^2(\pi\alpha)}{\pi^2\alpha^2 I_0^2(\pi\alpha)} e^{-N^2(\pi\alpha \coth(\pi\alpha) - 1)\Delta\theta^2/(\pi^2\alpha^2)}, \quad (\text{D5})$$

where we have replaced  $\hat{\theta} - \theta = \Delta\theta$ . Taking the integral over  $\Delta\theta$  then gives

$$\frac{\sinh^2(\pi\alpha)}{NI_0^2(\pi\alpha)\sqrt{\pi\alpha}\sqrt{\pi\alpha \coth(\pi\alpha) - 1}}. \quad (\text{D6})$$

It is found that this expression is asymptotically

$$\frac{\pi}{2N\sqrt{\alpha}} + \mathcal{O}(\alpha^{-3/2}). \quad (\text{D7})$$

This can be found using the asymptotic properties of Bessel functions,

$$\frac{1}{I_0^2(\pi\alpha)} \approx \frac{2\pi^2\alpha}{e^{2\pi\alpha}} \quad (\text{D8})$$

and  $\sinh^2(\pi\alpha) \approx e^{2\pi\alpha}/4$  and  $\coth(\pi\alpha) \approx 1$ , so

$$\begin{aligned} & \frac{\sinh^2(\pi\alpha)}{NI_0^2(\pi\alpha)\sqrt{\pi\alpha}\sqrt{\pi\alpha \coth(\pi\alpha) - 1}} \\ & \approx \frac{2\pi^2\alpha}{e^{2\pi\alpha}} \frac{e^{2\pi\alpha}}{4} \frac{1}{N\sqrt{\pi\alpha}\sqrt{\pi\alpha - 1}}. \end{aligned} \quad (\text{D9})$$

That gives the asymptotic expression claimed.

Now, for the integral over the tails we can upper bound the probability by that where we replace the sin with 1, so we have an upper bound

$$\begin{aligned} & 2 \int_{(\pi/N)\sqrt{1+\alpha^2}}^{\infty} \frac{1}{I_0^2(\pi\alpha)[N^2\delta\theta^2 - (\pi\alpha)^2]} d\Delta\theta \\ & = \frac{1}{I_0^2(\pi\alpha)} \frac{2 \operatorname{arcsinh}(\alpha)}{\pi N\alpha}. \end{aligned} \quad (\text{D10})$$

Now using  $\operatorname{arcsinh}(\alpha) \approx \ln(2\alpha)$ , we have the asymptotic expression

$$\frac{2\pi^2\alpha}{e^{2\pi\alpha}} \frac{2 \ln(2\alpha)}{\pi N\alpha} = \frac{4\pi \ln(2\alpha)}{Ne^{2\pi\alpha}}. \quad (\text{D11})$$

Dividing by the asymptotic expression for the normalization then gives

$$\frac{4\pi \ln(2\alpha)}{Ne^{2\pi\alpha}} \frac{2N\sqrt{\alpha}}{\pi} = 8 \ln(2\alpha)\sqrt{\alpha} e^{-2\pi\alpha}. \quad (\text{D12})$$

This tells us that, if we want probability of error outside the range given by  $\delta$ , then we should take

$$\ln(1/\delta) \approx 2\pi\alpha - \ln[8 \ln(2\alpha)\sqrt{\alpha}]. \quad (\text{D13})$$

Solving for  $\alpha$  then gives

$$\alpha = (1/2\pi) \ln(1/\delta) + \mathcal{O}(\ln \ln(1/\delta)). \quad (\text{D14})$$

The size of the confidence interval is  $(\pi/N)\sqrt{1+\alpha^2}$ , so if that needs to be  $\epsilon$ , we should take

$$N = \frac{\pi}{\epsilon} \sqrt{1+\alpha^2} = \frac{1}{2\epsilon} \ln(1/\delta) + \mathcal{O}(\epsilon^{-1} \ln \ln(1/\delta)). \quad (\text{D15})$$

The higher-order  $\ln \ln$  term for  $\alpha$  is larger than the correction term for approximating  $\sqrt{1+\alpha^2}$  by  $\alpha$ . The number of calls to  $U$  or  $U^\dagger$  is  $N$ , giving the complexity stated.

This derivation is for the continuous window, but the performance of the discrete window must be at least as good. Let  $f(m)$  be the continuous form of the window function, here the Kaiser window. Then take the Fourier transform to give  $\tilde{f}(\Delta\theta)$ , and define  $\tilde{g}(\Delta\theta)$  to be equal to  $\tilde{f}(\Delta\theta)$ , except zero for  $|\Delta\theta| > R$  with confidence interval half-width  $R$ ; that is

$$\tilde{g}(\Delta\theta) = \begin{cases} \tilde{f}(\Delta\theta), & |\Delta\theta| \leq R \\ 0, & |\Delta\theta| > R. \end{cases} \quad (\text{D16})$$

For tail probability  $\delta$  for  $\tilde{f}(\Delta\theta)$ , we then have

$$\begin{aligned} \delta &= \int d(\Delta\theta) \left| \tilde{f}'(\Delta\theta) - \tilde{g}'(\Delta\theta) \right|^2 \\ &= \int dm |f(m) - g(m)|^2, \end{aligned} \quad (\text{D17})$$

where the second equality is from Parseval's theorem.

That is for the continuous case, where the Fourier transform is not periodic in  $\Delta\theta$ . For the discrete case the Fourier transform is equal to the original Fourier transform convolved with a comb function with spacing  $2\pi$ . Let us use subscript  $d$  to indicate the discrete Fourier transforms of  $f$  and  $g$ . Since  $\tilde{g}(\Delta\theta)$  is zero for  $|\Delta\theta| > R$ , there is no overlap coming from the periodic images, and  $\tilde{g}_d$  is still zero for  $|\Delta\theta| > R$  (in  $[-\pi, \pi]$ ). The tail probabilities for  $\tilde{f}_d(\Delta\theta)$  can therefore be upper bounded by

$$\begin{aligned} &\int_{-\pi}^{-R} d(\Delta\theta) \left| \tilde{f}_d(\Delta\theta) \right|^2 + \int_R^{\pi} d(\Delta\theta) \left| \tilde{f}_d(\Delta\theta) \right|^2 \\ &\leq \int_{-\pi}^{\pi} d(\Delta\theta) \left| \tilde{f}_d(\Delta\theta) - \tilde{g}_d(\Delta\theta) \right|^2 \\ &= \sum_m |f(m) - g(m)|^2, \end{aligned} \quad (\text{D18})$$

where the equality is again from Parseval's theorem.

It is also possible to choose the discrete form of the window by shifting the sampling points by  $\nu \in [-1/2, 1/2]$ , which would give

$$\sum_m |f(m + \nu) - g(m + \nu)|^2. \quad (\text{D19})$$

Averaging over  $\nu$  then gives

$$\begin{aligned} &\int_{-1/2}^{1/2} d\nu \sum_m |f(m + \nu) - g(m + \nu)|^2 \\ &= \int dm |f(m) - g(m)|^2 = \delta. \end{aligned} \quad (\text{D20})$$

Therefore, the upper bound on the tail probability averaged over the shift  $\nu$  is equal to the tail probability in the continuous case,  $\delta$ . It is therefore possible to obtain performance

at least as good as the continuous case by an appropriate selection of the discrete window.

## APPENDIX E: QUBITIZATION WITH PROJECTION

Here we derive the expression for qubitization with a more general projection as given in Eq. (37). When the block encoding is defined more generally using

$$(|0\rangle\langle 0| \otimes P) V (|0\rangle\langle 0| \otimes P) = |0\rangle\langle 0| \otimes H/\lambda, \quad (\text{E1})$$

then for  $|k\rangle$  an eigenstate of  $H$  with energy  $E_k$  (and  $P|k\rangle = |k\rangle$ ), we have

$$\begin{aligned} (|0\rangle\langle 0| \otimes P) V (|0\rangle\langle 0| \otimes P) |0\rangle|k\rangle &= \frac{E_k}{\lambda} |0\rangle|k\rangle \\ (|0\rangle\langle 0| \otimes P) V |0\rangle|k\rangle &= \frac{E_k}{\lambda} |0\rangle|k\rangle \\ V |0\rangle|k\rangle &= \frac{E_k}{\lambda} |0\rangle|k\rangle + i\sqrt{1 - \left|\frac{E_k}{\lambda}\right|^2} |0k^\perp\rangle, \end{aligned} \quad (\text{E2})$$

where  $|0k^\perp\rangle$  is defined as a state such that  $(|0\rangle\langle 0| \otimes P)|0k^\perp\rangle = 0$ . That is,  $V$  gives an application of  $H$  to the target system, with failure being flagged by states orthogonal to  $|0\rangle$  on the ancilla or perpendicular to the projector  $P$  on the system. The phase factor on the orthogonal part can be chosen arbitrarily, and is chosen here as  $i$  to simplify later expressions. Then, one can define the qubitization as  $W := RV$  in the usual way except with the reflection being

$$R := i(2|0\rangle\langle 0| \otimes P - I). \quad (\text{E3})$$

This is similar to that in Ref. [24], except we have included the projection  $P$  in the reflection operation. Essentially the entire chain of reasoning as in Ref. [24] can be used, except replacing  $|0\rangle\langle 0| \otimes I$  with  $|0\rangle\langle 0| \otimes P$ . Then we obtain

$$W |0\rangle|k\rangle = i\frac{E_k}{\lambda} |0\rangle|k\rangle + \sqrt{1 - \left|\frac{E_k}{\lambda}\right|^2} |0k^\perp\rangle. \quad (\text{E4})$$

Then, to show the correct expression for  $W|\chi k^\perp\rangle$  we use



$$\begin{aligned}
W^\dagger |\chi\rangle |k\rangle &= -VR |\chi\rangle |k\rangle = -iV |\chi\rangle |k\rangle = iRW |\chi\rangle |k\rangle = iR \left( i\frac{E_k}{\lambda} |\chi\rangle |k\rangle + \sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} |\chi k^\perp\rangle \right) \\
&= -i\frac{E_k}{\lambda} |\chi\rangle |k\rangle + \sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} |\chi k^\perp\rangle. \tag{E5}
\end{aligned}$$

Similarly, for  $|\chi k^\perp\rangle$ , we have

$$W^\dagger |\chi k^\perp\rangle = -VR |\chi k^\perp\rangle = iV |\chi k^\perp\rangle. \tag{E6}$$

Now applying  $W^\dagger$  to the expression for  $W |\chi\rangle |k\rangle$  gives

$$\begin{aligned}
|\chi\rangle |k\rangle &= i\frac{E_k}{\lambda} W^\dagger |\chi\rangle |k\rangle + \sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} W^\dagger |\chi k^\perp\rangle \\
&= i\frac{E_k}{\lambda} \left( -i\frac{E_k}{\lambda} |\chi\rangle |k\rangle + \sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} |\chi k^\perp\rangle \right) + i\sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} V |\chi k^\perp\rangle \\
&= \left| \frac{E_k}{\lambda} \right|^2 |\chi\rangle |k\rangle + i\frac{E_k}{\lambda} \sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} |\chi k^\perp\rangle + i\sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} V |\chi k^\perp\rangle \\
\left( 1 - \left| \frac{E_k}{\lambda} \right|^2 \right) |\chi\rangle |k\rangle &= i\frac{E_k}{\lambda} \sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} |\chi k^\perp\rangle + i\sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} V |\chi k^\perp\rangle \\
-i\sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} |\chi\rangle |k\rangle &= \frac{E_k}{\lambda} |\chi k^\perp\rangle + V |\chi k^\perp\rangle \\
V |\chi k^\perp\rangle &= -\frac{E_k}{\lambda} |\chi k^\perp\rangle - i\sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} |\chi\rangle |k\rangle \\
RV |\chi k^\perp\rangle &= i\frac{E_k}{\lambda} |\chi k^\perp\rangle + \sqrt{1 - \left| \frac{E_k}{\lambda} \right|^2} |\chi\rangle |k\rangle. \tag{E7}
\end{aligned}$$

Hence we obtain Eq. (37) as required. We have corrected the extra factors of  $i$  included in Ref. [24]. Note that the rest of the reasoning in Ref. [24] is correct.

## APPENDIX F: BETTI NUMBER AND SPECTRAL GAP CALCULATIONS

The purpose of this section is to prove Propositions 1, 2, 3, and 4. In this section, we will work with *reduced* homology. This is identical to regular homology, except that we have an extra one-dimensional space  $C_{-1}$  and an extra boundary map  $\partial_0 : C_0 \rightarrow C_{-1}$ , which maps every vertex (0-simplex) to the unique basis vector of  $C_{-1}$ . This has the effect that the reduced homology  $H_0$  is equal to the number of connected components *minus one*, rather than simply the number of connected components. The rest of the homology groups  $H_k$  for  $k > 0$  are unchanged.

*Definition 9.*—Given two simplicial complexes  $X$  and  $Y$ , define their *join*  $X * Y$  to be the simplicial complex consisting of faces  $\sigma \otimes \tau := \sigma \cup \tau$  for all  $\sigma \in X, \tau \in Y$ .

We observe that  $K(m, k) = K(m, k-1) * K(m, 1)$ . Moreover, the homology of the join is given by the well-known Kunneth formula.

*Lemma 7.*—(Kunneth formula)

$$\tilde{H}_k(X * Y) = \bigoplus_{i+j=k-1} \tilde{H}_i(X) \otimes \tilde{H}_j(Y) \tag{F1}$$

$$\implies \tilde{\beta}_k(X * Y) = \sum_{i+j=k-1} \tilde{\beta}_i(X) \tilde{\beta}_j(Y). \tag{F2}$$

We would also like to relate the Laplacian of  $X * Y$  to the Laplacians of  $X$  and  $Y$ .

*Lemma 8.*—Let  $\sigma \in X$  be an  $i$ -simplex and  $\tau \in Y$  a  $j$ -simplex with  $i + j = k - 1$ . Then

$$\Delta_k^{X*Y}(\sigma \otimes \tau) = (\Delta_i^X \sigma) \otimes \tau + \sigma \otimes (\Delta_j^Y \tau). \quad (\text{F3})$$

*Proof.*—Let us work in the graded algebra  $C_{-1} \oplus C_0 \oplus C_1 \oplus \dots$ . We have

$$\begin{aligned} \Delta &= \partial^\dagger \partial + \partial \partial^\dagger \\ \partial(\sigma \otimes \tau) &= (\partial\sigma) \otimes \tau + (-1)^{|\sigma|} \sigma \otimes (\partial\tau) \\ \partial^\dagger(\sigma \otimes \tau) &= (\partial^\dagger\sigma) \otimes \tau + (-1)^{|\sigma|} \sigma \otimes (\partial^\dagger\tau) \\ \implies \Delta(\sigma \otimes \tau) &= (\Delta\sigma) \otimes \tau + \sigma \otimes (\Delta\tau) \quad \blacksquare \end{aligned}$$

*Corollary 1.*—Let  $\text{spec } \Delta$  denote the set of eigenvalues of  $\Delta$ .

$$\text{spec } \Delta_k^{X*Y} = \bigcup_{i+j=k-1} \text{spec } \Delta_i^X + \text{spec } \Delta_j^Y. \quad (\text{F4})$$

Here the plus notation for sets means  $A + B = \{a + b : a \in A, b \in B\}$ .

*Proof.*—Use Lemma 8 and let  $\sigma \in C_i^X$  and  $\tau \in C_j^Y$  be eigenchains of  $\Delta_i^X$  and  $\Delta_j^Y$ , respectively.  $\blacksquare$

*Proposition 6.*—(Restatement of Proposition 1.)

The  $(k - 1)$ th Betti number of (the clique complex of)  $K(m, k)$  is

$$\beta_{k-1} = (m - 1)^k. \quad (\text{F5})$$

*Proof.*— $K(m, k) = K(m, k - 1) * K(m, 1)$  and the Betti numbers of  $K(m, 1)$  are  $(m - 1, 0, 0, \dots)$ . Thus by induction using the Kunneth formula, we have  $\beta_{k-1} = (m - 1)^k$ .  $\blacksquare$

*Proposition 7.*—(Restatement of Proposition 2.)

The combinatorial Laplacian  $\Delta_{k-1} = \partial_{k-1}^\dagger \partial_{k-1} + \partial_k \partial_k^\dagger$  of (the clique complex of)  $K(m, k)$  has spectral gap

$$\lambda_{\min} = m. \quad (\text{F6})$$

*Proof.*—Again  $K(m, k) = K(m, k - 1) * K(m, 1)$ . The spectrum of the  $\Delta_0^{K(m, 1)}$  is 0 with multiplicity  $m - 1$ , and  $m$  with multiplicity 1. Thus by induction using Corollary 1, the spectrum of  $\Delta_{k-1}^{K(m, k)}$  is (ignoring multiplicities)  $\{0, m, 2m, \dots, km\}$ . This gives  $\lambda_{\min} = m$ .  $\blacksquare$

*Proposition 8.*—(Restatement of Proposition 3.)

The Rips complex described in Sec. IV C has  $(2k - 1)$ th Betti number given by

$$\beta_{2k-1}(S) = (m - 1)^k = \left(\frac{n}{2k} - 1\right)^k. \quad (\text{F7})$$

*Proof.*—The Betti numbers of  $\mathcal{R}_1(S_i)$  are  $(m - 1, 0, \dots)$ . Furthermore,  $R_1(S) = R_1(S_0) * \dots * R_1(S_{k-1})$ . Thus by the Kunneth formula  $\beta_{2k-1}(S) = (m - 1)^k$ .  $\blacksquare$

*Proposition 9.*—(Restatement of Proposition 4.)

The Rips complex described in Sec. IV C has a combinatorial Laplacian  $\Delta_k = \partial_k^\dagger \partial_k + \partial_{k+1} \partial_{k+1}^\dagger$  with a constant spectral gap  $\lambda_{\min}$ .

*Proof.*—Consider first the Laplacian  $\Delta_k^{S_i}$  of  $\mathcal{R}_1(S_i)$ . Say it has smallest eigenvalue  $c$ . Then by induction using Corollary 1, the smallest eigenvalue of  $\Delta_k^S$  is also  $c$ .  $\blacksquare$

## 1. Perturbations of $K(m, k)$ and smaller spectral gaps

In Sec. IV D, we introduced a novel classical algorithm that, in conjunction with Apers *et al.*'s algorithm [36], significantly enhances classical runtimes in the domain where quantum algorithms achieve polynomial runtimes. Moreover, these classical algorithms can potentially operate in polynomial time, contingent on the required precision for estimating the Betti number, provided the spectral gap of the combinatorial Laplacian is sufficiently large. In this section, we study perturbations of the graph  $K(m, k)$  (defined in the previous section) that lead to a smaller spectral gap. These perturbations render the classical algorithms inefficient, regardless of the precision needed for the Betti-number estimation, though they do not impair the efficiency of the quantum algorithm. Additionally, these perturbations also show that the speedups are robust, in the sense of a much larger class of graphs having a guaranteed speedup.

We will denote our perturbed family of graphs  $K'(m, k)$ . Recall that  $K(m, k)$  consisted of  $k$  clusters, each with  $m$  vertices. There were no edges within the clusters, and all edges between vertices in different clusters were included. Let  $G'$  be the graph consisting of  $m$  vertices with a single edge. Thus  $G'$  consists of a single edge, and  $m - 2$  additional completely disconnected vertices. Let  $K'(m, k)$  consist of  $k$  clusters, where each cluster is a copy of  $G'$ . Likewise, we include all edges between clusters. In other words, we obtain an instance of  $K'$  by adding a single edge to each of the clusters, which were originally independent sets.

*Lemma 9.*—The  $(k - 1)$ th Betti number of the clique complex of  $K'(m, k)$  is

$$\beta_{k-1} = (m - 2)^k.$$

*Lemma 10.*—The  $(k - 1)$ th combinatorial Laplacian  $\Delta_{k-1}$  of the clique complex of  $K'(m, k)$  has smallest nonzero eigenvalue

$$\lambda_{\min} = 2,$$

and largest eigenvalue

$$\lambda_{\max} = 2mk = 2n.$$

*Proof.*—First, note that the clique complex of  $K(m, k)$  is the  $k$ -fold join of the clique complex of the graph  $G'$  consisting of  $m$  vertices with a single edge between two of the vertices. Next, use the Kunneth formula and Corollary 1. Finally, note that the zeroth combinatorial Laplacian of  $\text{Cl}(G')$  is

$$\Delta_0 = \begin{pmatrix} 2 & 0 & 1 & 1 & \dots \\ 0 & 2 & 1 & 1 & \dots \\ 1 & 1 & 1 & 1 & \dots \\ 1 & 1 & 1 & 1 & \dots \\ \cdot & \cdot & \cdot & \cdot & \dots \end{pmatrix}.$$

The spectrum of  $\Delta_0$  is thus 0 with multiplicity  $m - 2$ , 2 with multiplicity  $m$ , and  $2m$  with multiplicity 1. ■

*Corollary 2.*—The  $(k - 1)$ th combinatorial Laplacian  $\Delta_{k-1}$  of the clique complex of  $K'(n/k, k)$  has normalized spectral gap

$$\gamma = \frac{\lambda_{\min}}{\lambda_{\max}} = \frac{1}{n}.$$

In short, when  $k = c \times \log(n)$  the graph  $K'(m, k)$  has an inverse polynomial spectral gap, as opposed to an inverse logarithmic spectral gap as in the case for  $K(m, k)$ . This key distinction amplifies the disparities in runtime between the quantum and classical algorithms. Notably, the quantum algorithm (still) attains a superpolynomial speedup over its classical counterparts, even in scenarios where only a constant level of additive precision for the normalized Betti number is required.

Next, we demonstrate the potential for even broader generalization of these perturbations. Consider the  $m$ -vertex graph  $G''$  that is a disjoint union of  $m/2$  edges (i.e., there are no isolated vertices), and let  $K''(m, k)$  be the  $k$ -fold join of  $G''$ .

*Lemma 11.*—The spectrum of the  $(k - 1)$ th combinatorial Laplacian of the clique complex of  $K''(m, k)$  is

$$\text{spec } \Delta_{k-1} = \left\{ \lambda_{i,j,\ell} \mid i + j + \ell = k \right\},$$

where

$$\lambda_{i,j,\ell} = 0 \times i + 2 \times j + m \times \ell$$

and they have multiplicities

$$\text{multiplicity}[i, j, \ell] = \binom{n}{i, j, \ell} \times \left(\frac{m}{2} - 1\right)^i \times \left(\frac{m}{2}\right)^j \times 1^\ell.$$

*Proof.*—The 0th combinatorial Laplacian of  $G''$  is

$$\Delta_0(\text{Cl}(G'')) = \begin{pmatrix} 2 & 0 & 1 & 1 & \dots \\ 0 & 2 & 1 & 1 & \dots \\ 1 & 1 & 2 & 0 & \dots \\ 1 & 1 & 0 & 2 & \dots \\ \cdot & \cdot & \cdot & \cdot & \dots \end{pmatrix}.$$

A quick computation reveals that  $\Delta_0$  has a spectrum

$$\begin{aligned} &\text{spec}(\Delta_0(\text{Cl}(G''))) \\ &= \{0 \text{ (with multiplicity } m/2 - 1), \\ &\quad 2 \text{ (with multiplicity } m/2), m \text{ (with multiplicity } 1)\}. \end{aligned}$$

Following this observation, the lemma follows from the application of Corollary 1. ■

*Corollary 3.*—The  $(k - 1)$ th combinatorial Laplacian  $\Delta_{k-1}$  of the clique complex of  $K''(n/k, k)$  has normalized spectral gap

$$\gamma = \frac{\lambda_{\min}}{\lambda_{\max}} = \frac{2}{n}.$$

Again, the graph  $K''(m, k)$  has an inverse polynomial spectral gap, as opposed to an inverse logarithmic spectral gap. In particular, the quantum algorithm (still) attains a superpolynomial speedup over its classical counterparts, even in scenarios where only a constant level of additive precision for the normalized Betti number is required. In conclusion, adding a single edge or  $m/2$  edges to each cluster within  $K(m, k)$  does not change the superpolynomial quantum speedup that it exhibits.

## APPENDIX G: DEQUANTIZATION USING PATH-INTEGRAL MONTE CARLO

Previously, we argued that the cases where TDA can potentially have a superpolynomial advantage relative to classical approaches is in cases where the clique density is high. In such cases sampling in our quantum algorithm is efficient and eigendecomposition is inefficient. However, we will see here that this is not necessarily the case and that there are cases where the clique density is high wherein randomized classical algorithms can achieve scaling that is polynomially equivalent to quantum algorithms. This will show that the conditions for a substantial improvement using quantum TDA are potentially even more subtle than previous work suggests.

As discussed in the main body, our dequantization looks at imaginary-time simulations of the Hermitian operator  $\widetilde{B}_G^2 = B_G^2 + (1 - P)\gamma_{\min}$ , where  $B_G$  is the square of the constrained Dirac operator, i.e., the combinatorial Laplacian, and the projector selects all input states that are valid simplices. Given an upper bound on the sparsity  $s$  of the combinatorial Laplacian, we can obtain the decomposition  $\widetilde{B}_G^2 = \sum_{p=1}^D c_p H_p$ , where  $H_p$  is one sparse, Hermitian and unitary, and  $D = \mathcal{O}(s^2)$ , in polynomial time using distributed graph coloring algorithms. These algorithms also let us compute the position of the nonzero matrix element in row  $x$  of  $U_\alpha$  using a number of queries to  $\widetilde{B}_G^2$  that scales as  $\mathcal{O}(\log^*(D))$  [45]. We will provide explicit bounds on the sparsity of the combinatorial Laplacian in a subsequent section.

In order to set up the relevant path integrals, we must first employ a Trotter decomposition. This allows us to represent the exponential in terms of exponentials of the one-sparse matrices, which can then be simulated through randomization. This leads us to the conclusion that

$$e^{-\tilde{B}_G^2 t} = (e^{-\tilde{B}_G^2 t/r})^r = \left( \prod_{p=1}^D e^{-c_p H_p t/2r} \prod_{p=D}^1 e^{-c_p H_p t/2r} + \mathcal{O}\left(\frac{(\sum_p |c_p|)^3 t^3}{r^3}\right) \right)^r. \quad (\text{G1})$$

As  $H_p$  is Hermitian, it has a complete set of eigenvectors. Since  $H_p$  is also one sparse and as such matrices can be written as the direct sum of irreducible one- and two-dimensional matrices, we can parameterize the eigenvectors to respect the structure of the two-dimensional space via

$$H_p |\lambda_{p,v}\rangle = \lambda_{p,v} |\lambda_{p,v}\rangle. \quad (\text{G2})$$

Note that each eigenvector  $|\lambda_{p,v}\rangle$  is such that  $\langle p | \lambda_{p,v}\rangle$  is nonzero for only two different computational basis vectors.

Note that the first term on the rhs of Eq. (G1) contains  $2rD$  terms. If we introduce a vector of indices  $p = \{1, \dots, r, 1, \dots, r, \dots, r\}$  with  $2rD$  entries denoted by  $p_i$ , we can express the expectation of the exponential of the boundary operator as ( $H$  stands for the Haar average)

$$\mathbb{E}_{|\psi\rangle}^H \left( \prod_{i=1}^{2rD} e^{-c_{p_i} H_{p_i} t/2r} \right) := \mathbb{E}_H \langle \psi | \left( \prod_{i=1}^{2rD} e^{-c_{p_i} H_{p_i} t/2r} \right) | \psi \rangle. \quad (\text{G3})$$

Next we set up our path integrals by selecting sets of  $2rD$  indices that correspond to the eigenstates that we transition to in the path integral. We denote such a path via the vector  $\Gamma$  where  $\Gamma_j$  corresponds to the index of the  $j$ th eigenstate in the path. Using this notation we can insert resolutions of the identity of the form  $\sum_{\Gamma_j} |\lambda_{p_j, \Gamma_j}\rangle \langle \lambda_{p_j, \Gamma_j}|$  consisting of the eigenvectors  $|\lambda_{p_j, \Gamma_j}\rangle$  of each  $U_{p_j}$  in between each of the  $2rD$  terms and defining  $\rho = |\psi\rangle \langle \psi|$  gives

$$\begin{aligned} \mathbb{E}_{|\psi\rangle}^H \left( \prod_{i=1}^{2rD} e^{-c_{p_i} H_{p_i} t/2r} \right) &= \mathbb{E}_H \text{Tr} \left( \rho \prod_{i=1}^{2rD} e^{-c_{p_i} H_{p_i} t/2r} \right) \\ &= \mathbb{E}_H \text{Tr} \left( \rho \sum_{\Gamma_1, \dots, \Gamma_{2rD}} \exp \left( - \sum_{i=1}^{2rD} \lambda_{p_i, \Gamma_i} t/2r \right) |\lambda_{p_1, \Gamma_1}\rangle \langle \lambda_{p_1, \Gamma_1}| \cdots |\lambda_{p_{2rD}, \Gamma_{2rD}}\rangle \langle \lambda_{p_{2rD}, \Gamma_{2rD}}| \right) \\ &= \mathbb{E}_H \text{Tr} \left( \rho \sum_{\Gamma_1, \dots, \Gamma_{2rD}} \exp \left( - \sum_{i=1}^{2rD} \lambda_{p_i, \Gamma_i} t/2r \right) W(\Gamma) |\lambda_{p_1, \Gamma_1}\rangle \langle \lambda_{p_{2rD}, \Gamma_{2rD}}| \right) \\ &= \mathbb{E}_H \text{Tr} \left( \rho \mathbb{E}_\Gamma \frac{\exp \left( - \sum_{i=1}^{2rD} \lambda_{p_i, \Gamma_i} t/2r \right) W(\Gamma) |\lambda_{p_1, k_1}\rangle \langle \lambda_{p_{2rD}, k_{2rD}}|}{\text{Pr}(\Gamma)} \right) \\ &= \mathbb{E}_H \mathbb{E}_\Gamma \text{Tr} \left( \rho \frac{\exp \left( - \sum_{i=1}^{2rD} \lambda_{p_i, \Gamma_i} t/2r \right) W(\Gamma) |\lambda_{p_1, \Gamma_1}\rangle \langle \lambda_{p_{2rD}, \Gamma_{2rD}}|}{\text{Pr}(\Gamma)} \right) \\ &= \frac{1}{d_{k-1}} \mathbb{E}_\Gamma \text{Tr} \left( \frac{\exp \left( - \sum_{i=1}^{2rD} \lambda_{p_i, \Gamma_i} t/2r \right) W(\Gamma) |\lambda_{p_1, \Gamma_1}\rangle \langle \lambda_{p_{2rD}, \Gamma_{2rD}}|}{\text{Pr}(\Gamma)} \right) \\ &= \frac{1}{d_{k-1}} \mathbb{E}_\Gamma \left( \frac{\exp \left( - \lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/2r \right) W(\Gamma) \delta_{\Gamma_1, \Gamma_{2rD}}}{\text{Pr}(\Gamma)} \right), \end{aligned} \quad (\text{G4})$$

where we have defined for convenience the quantity

$$W(\Gamma) = \langle \lambda_{p_1, \Gamma_1} | \lambda_{p_2, \Gamma_2} \rangle \cdots \langle \lambda_{p_{2rD-1}, \Gamma_{2rD-1}} | \lambda_{p_{2rD}, \Gamma_{2rD}} \rangle, \quad (\text{G5})$$

with  $\Gamma = [\Gamma_1, \dots, \Gamma_{2rD}]$ . In the fourth line, we divided and multiplied by a probability  $\text{Pr}(\Gamma)$  to express the sum as an average, which allows us to use importance sampling to minimize the variance via a judicious choice of  $\text{Pr}(\Gamma)$ . In the last line, we used the fact that  $p_{2rD} = p_1$  for the symmetric Trotter formula.

If we wish to estimate this value by sampling, the primary driver of the complexity will be the estimation of the expectation value through the sample mean, which corresponds to the optimal unbiased estimator of the population mean. The number of samples scales with the variance of the set that one averages over and the variance over  $\Gamma$  of the above Haar expectation is then simply

$$\begin{aligned}
& \mathbb{V}_\Gamma \left( \frac{1}{d_{k-1}} \text{Tr} \left( \frac{\exp \left( -\sum_{i=1} \lambda_{p_i, \Gamma_i} t/2r \right) W(\Gamma) |\lambda_{p_1, \Gamma_1}\rangle \langle \lambda_{p_{2rD}, \Gamma_{2rD}}|}{\text{Pr}(\Gamma)} \right) \right) \\
&= \frac{1}{d_{k-1}^2} \sum_{\Gamma_1, \dots, \Gamma_{2rD-1}} \frac{\exp \left( -2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r \right) |W(\Gamma)|^2 \delta_{\Gamma_1, \Gamma_{2rD}}}{\text{Pr}(\Gamma)} \\
&\quad - \left( \frac{1}{d_{k-1}} \mathbb{E}_\Gamma \left( \frac{\exp \left( -\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/2r \right) W(\Gamma) \delta_{\Gamma_1, \Gamma_{2rD}}}{\text{Pr}(\Gamma)} \right) \right)^2 \\
&\leq \frac{1}{d_{k-1}^2} \sum_{\Gamma_1, \dots, \Gamma_{2rD-1}} \frac{\exp \left( -2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r \right) |W(\Gamma)|^2 \delta_{\Gamma_1, \Gamma_{2rD}}}{\text{Pr}(\Gamma)}. \tag{G6}
\end{aligned}$$

There are many probability distributions that we could choose to sample from to minimize the variance in Eq. (G6). The most straightforward distribution to choose, and the appropriate one to pick in the limit of short  $t$ , is a uniform distribution. However, in practice the eigenvalues in the sum may have wildly varying sizes and so the importance of each of the different paths can swing substantially. A more natural choice to make for the probability of drawing each path is

$$\text{Pr}(\Gamma) = \frac{\exp \left( -2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r \right) \delta_{\Gamma \in S_\Gamma}}{\sum_{\Gamma \in S_\Gamma} \exp \left( -2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r \right)}, \tag{G7}$$

where  $S_\Gamma$  is the set of valid paths with  $2rD$  vertices such that each edge corresponds to a path of connected eigenvectors for the one-sparse matrices used in the decomposition of  $\widetilde{B}_G^2$ .

The central challenge in employing this formula is to estimate the value of the sums over the values of  $\Gamma_i$ . As the  $H_j$  in used in the derivation are Hermitian and unitary, we have that

$$\lambda_{p_i, \Gamma_i} = \pm c_{p_i}. \tag{G8}$$

Furthermore, the one-sparse decomposition is chosen in such a way that the matrix elements are all off diagonal with the exception of any diagonal matrix that appears in the decomposition. This can be seen explicitly using

the discussion of the Jordan-Wigner representation of the Dirac operator. This means that each eigenvector couples to at most two eigenvectors. At most one term is diagonal in the standard Trotter decomposition of the Dirac operator [45]. A simple combinatorial argument therefore leads to the conclusion that the total number of valid paths is at most  $d_{k-1} 2^{2r(D-1)-1}$ .

Given this choice, the normalization constant (which is analogous to a partition function) can be expressed (assuming that the diagonal element is always  $p_i = D$ ) as

$$\begin{aligned}
& \sum_{\Gamma \in G} \exp \left( -2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r \right) \\
&= d_{k-1} 2^{2rD-1-r} \cosh(2c_{p_1} t/r) \prod_{i=2}^{2rD-1} (\cosh(c_{p_i} t/r) \delta_{p_i \neq D} \\
&\quad + \delta_{p_i = D} e^{-\lambda_{p_i, \Gamma_i} t/r} / 2). \tag{G9}
\end{aligned}$$

This can be computed using  $\mathcal{O}(\text{poly}(rD))$  arithmetic operations and so does not ruin the efficiency of the algorithm. Note that were the sum over the  $W(\Gamma)$  terms considered instead, then the result would be computationally difficult to compute as these terms generate correlations that would prevent us from performing an independent sum for each of the factors.

The variance  $\sigma^2$  over the values of  $k$  chosen in the path integrals for the expression for the Haar average is then

$$\begin{aligned}
\sigma^2 &= \frac{1}{d_{k-1}^2} \left( \sum_{\Gamma \in S_\Gamma} |W(\Gamma)|^2 \right) \left( \sum_{\Gamma \in S_\Gamma} \exp \left( -2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r \right) \right) \\
&\quad - \frac{1}{d_{k-1}^2} \left| \sum_{\Gamma \in S_\Gamma} \exp \left( -\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/2r \right) W(\Gamma) \right|^2 \\
&\leq \frac{2^{2rD} e^{2Dt \max_i \lambda_{p_i, \Gamma_i}}}{d_{k-1}} \leq \frac{2^{2rD} (1/\delta)^{2D \max_i |\lambda_{p_i, \Gamma_i}|/\gamma_{\min}}}{d_{k-1}}. \tag{G10}
\end{aligned}$$

This shows that the variance after making this substitution is precisely equal to the gap in the Cauchy-Schwarz inequality in the latter sum. This suggests that in certain cases where the Cauchy-Schwarz inequality is tight, the variance may be extremely small given that we have the ability to sample from the distribution  $\Pr(\Gamma)$ .

### 1. Metropolis-Hastings algorithm

Outside of specific cases such as graphs, it is difficult in general to sample from the probability distribution  $\Pr(\Gamma)$  to employ the above variance reduction strategy. It is therefore necessary to provide a general method to obtain these samples if we wish to understand how we could address the problem more generally. One way to address the issue of how to sample from the distribution  $\Pr(\Gamma)$  is to use the Metropolis-Hastings algorithm. The idea behind the algorithm is to design a Markov chain whose stationary distribution equals our choice of  $\Pr(\Gamma)$ .

We first start with a connected, undirected graph on the set of all possible states  $\Gamma_1, \Gamma_2, \dots, \Gamma_{2rD-1}$  which represent the ‘‘paths’’ involved in our Trotter decomposition. Each vertex of the graph then represents one possible collection of values for  $\Gamma_1, \dots, \Gamma_{2rD-1}$ .

At each vertex  $a$ , we therefore select a neighbor with probability  $1/(2rD-1)$ . Since the degree may be less than  $2rD-1$  at a given vertex, the walk may remain at that vertex as there is a nonzero probability of no edge being selected. To account for such situations, we have the following rules: if a neighboring vertex  $b$  is selected and the probability of transitioning

$$p_b := \Pr(\Gamma)_b = \exp \left( -2\lambda_{p_1, \Gamma_1^{(b)}} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i^{(b)}} t/r \right) \tag{G11}$$

is at least as great as the probability of remaining  $p_a := \Pr(\Gamma)_a$ , we transition to  $b$ . If  $p_b < p_a$ , then we transition to  $b$  with probability  $p_b/p_a$ , where  $b \in \mathcal{N}(a)$  with  $\mathcal{N}(a)$  referring to the neighbors of  $a$  and

$$\frac{p_b}{p_a} = \frac{\exp \left( -2\lambda_{p_1, \Gamma_1^{(b)}} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i^{(b)}} t/r \right)}{\exp \left( -2\lambda_{p_1, \Gamma_1^{(a)}} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i^{(a)}} t/r \right)}. \tag{G12}$$

Otherwise, we remain at  $a$  with probability  $1 - p_b/p_a$ . Defining

$$p_{ab} := \frac{1}{R} \min \left( 1, \frac{p_b}{p_a} \right) \tag{G13}$$

and

$$p_{aa} := 1 - \sum_{b \neq a} p_{ab}, \tag{G14}$$

we can easily verify  $p_a p_{ab} = p_b p_{ba}$ . By the fundamental theorem of Markov chains [53], it follows that the stationary probabilities are  $p_a$  as needed.

The cost of computing the ratio  $p_b/p_a$  is  $\mathcal{O}(rD)$  arithmetic operations, which coincides with the cost of performing an update. The number of such updates needed to reach  $\delta$  error from the stationary distribution, where  $\delta$  is the total variational distance (TVD) from the stationary distribution desired, is

$$T^* \in \mathcal{O} \left( \frac{\log(1/\delta)}{\gamma_M} \right), \tag{G15}$$

where  $\gamma_M := 1 - \lambda_2$  is the eigenvalue gap of for the transition matrix  $p$ . This implies that the number of arithmetic operations needed to sample from a distribution that is  $\delta$  close to the stationary distribution is in

$$\mathcal{O} \left( \frac{rD \log(1/\delta)}{\gamma_M} \right). \tag{G16}$$

If one samples from a distribution,  $P'$ , that is  $\delta$  close to the intended distribution  $P$  then the expectation value of any function  $f$  is  $|\sum_j (P(j)f(j)) - \sum_j (P'(j)f(j))| \leq \delta \max |f(j)|$ . Similarly, the variance obeys  $|\sum_j P'(j)f(j)^2 - (\sum_j P'(j)f(j))^2| \leq \mathbb{V}(f) + \mathcal{O}(\delta \max |f(j)|^2)$ . Thus if we want the error in the mean to be less than some error  $\epsilon_M$ , we require

$$\delta = \epsilon_M / \max |f(j)|. \quad (\text{G17})$$

Note that in our situation

$$|f| = \left| \frac{\exp\left(-\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/2r\right) W(\Gamma)}{\text{Pr}(\Gamma)} \right|, \quad (\text{G18})$$

where  $r$  is the number time steps needed in the Trotterization procedure to attain a desired Trotter error  $\epsilon_T$ . Further, we have the following bound:

$$\begin{aligned} & \sum_{\Gamma_1, \dots, \Gamma_{2rD-1}} \left| W(\Gamma) \exp\left(-2\lambda_{p_1, \Gamma_1} t/r - \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r\right) \right| \\ & \leq \sqrt{\sum_{\Gamma_1, \dots, \Gamma_{2rD-1}} |W(\Gamma)|^2} \sqrt{\sum_{\Gamma \in \mathcal{S}_\Gamma} \exp\left(-4\lambda_{p_1, \Gamma_1} t/r - 2 \sum_{i=2}^{2rD-1} \lambda_{p_i, \Gamma_i} t/r\right)}. \end{aligned} \quad (\text{G19})$$

Note that when we sum over a specific  $p_i, \Gamma_i$  in  $|W(\Gamma)|^2$ , we get 1 since the eigenvectors are orthonormal. Performing all  $2rD - 2$  sums therefore gives 1 from all the inner products and the last sum involving the  $(2rD - 1)$ th index gives a factor of  $d_{k-1}$ . Additionally, each  $\lambda_{p_i, \Gamma_i}$  can either be positive or negative but is upper bounded by  $\|B_G^2\|_{\max} \leq \|B_G^2\|_{\infty} = \gamma_{\max}$ , i.e., the largest eigenvalue of  $B_G^2$ . We can then bound  $f_{\max}$  as follows:

$$|f| \leq f_{\max} \leq d_{k-1} e^{2\gamma_{\max} t D} 2^{r(D-1/2)}. \quad (\text{G20})$$

## 2. Trotter error in path integration

From Corollary 12 of Ref. [54], the multiplicative Trotter error  $m$  for a  $p$ th-order Trotter formula is asymptotically bounded by

$$\mathcal{O}\left(\alpha \left(\frac{t}{r}\right)^{p+1} \exp\left(\frac{2t}{r} \Upsilon \sum_{\ell=1}^{\Gamma} \|H_{\ell}\|\right)\right), \quad (\text{G21})$$

where the operator  $H$  is decomposed into a sum of  $\Gamma$  terms,

$$\alpha = \sum_{\ell_1, \ell_2, \dots, \ell_{p+1}=1} \|[H_{\ell_{p+1}}, \dots, [H_{\ell_2}, H_{\ell_1}] \dots]\|, \quad (\text{G22})$$

and  $\Upsilon$  is the number of ‘‘stages’’ of the formula. For the symmetric Trotter-Suzuki formula,  $\Upsilon = 2(5)^{q-1}$  for a TS formula of order  $2q$ , where  $q = 1$  and  $p = 2$  for our case.  $\Gamma = 2rD$  and  $\alpha$  can be upper bounded by

$$\alpha \leq (2rD) 4 \max_p (c_p)^3 \leq 8rD \gamma_{\max}^3. \quad (\text{G23})$$

Since this is the short time multiplicative error bound for simulating an operator for time  $t/r$ , we want to bound the resulting error when simulating for large  $t$ . To this end, note that if we have an operator  $A$  we approximate by an operator  $B$  up to some multiplicative error  $m$ , then  $B = A(I + mC)$  where  $C$  is an operator such that  $\|C\| \leq 1$ , and  $m$  is a constant. Then

$$\begin{aligned} \|B\|^r & \leq (\|A\|(I + m\|C\|))^r \leq \|A\|^r \left(1 + \sum_{q=1}^r (m\|C\|)^q \binom{r}{q}\right) \\ & \leq \|A\|^r \left(1 + \sum_{q=1}^r m^q \binom{r}{q}\right) \leq \|A\|^r \left(1 + \sum_{q=1}^r \left(\frac{mre}{q}\right)^q\right) \\ & \leq \|A\|^r \left(1 + \sum_{q=1}^r (mre)^q\right) \leq \|A\|^r \left(1 + \frac{mre}{1 - mre}\right). \end{aligned} \quad (\text{G24})$$

Therefore, the long-time multiplicative error is bounded by  $mre/(1 - mre)$  and we would like this to be less than some desired error  $\epsilon_T > 0$ . This implies that we must have  $mre \leq \epsilon_T/(1 + \epsilon_T) \leq \epsilon_T$ . In our context,  $A = e^{tH/r}$ ,  $B$  is an approximation to  $A$  as given by a Trotter formula, and  $m$  is the short-time multiplicative Trotter error bound cited above.

Using the bound on  $m$  and substituting in the parameters relevant for our situation, we have

$$mre \leq e\alpha \frac{t^3}{r^2} \exp\left(\frac{4t}{r} \sum_{\ell=1}^{2rD} \|H_\ell\|\right) \leq \epsilon_T. \quad (\text{G25})$$

If  $r \geq \frac{4t}{\ln 2} \sum_{\ell} \|H_\ell\|$ , then

$$mre \leq 2e\alpha \frac{t^3}{r} \leq 2e\alpha \frac{t^3}{r^2} \leq \epsilon_T, \quad (\text{G26})$$

which implies that

$$r = t \max\left\{\left(\frac{4e\alpha}{\epsilon_T}\right)^{1/2}, \frac{4}{\ln 2} \sum_{\ell} \|H_\ell\|\right\}. \quad (\text{G27})$$

The former term dominates asymptotically, so we will henceforth take  $r \in \Theta\left(t(4e\alpha/\epsilon_T)^{1/2}\right)$ .

The systematic error in the estimate of the expectation value from the Trotter-Suzuki formula and the finite-length Markov chain is at most

$$\|e^{-\tilde{B}_G^2 t}\| \epsilon_T + \epsilon_M := \epsilon_{TM}. \quad (\text{G28})$$

The total number of operations needed to draw a single sample from the distribution with bias at most  $\epsilon_{TM}$  is from Eq. (G16) in

$$\mathcal{O}\left(\frac{rD \log(1/\delta)}{\gamma_M}\right) = \mathcal{O}\left(\frac{\sqrt{\alpha} t^{3/2} D \log(f_{\max}/\epsilon_M)}{\sqrt{\epsilon_T} \gamma_M}\right). \quad (\text{G29})$$

Next taking  $\epsilon_M = \epsilon_{TM}/2$  and similarly for  $\|e^{-\tilde{B}_G^2 t}\| \epsilon_T$ , we have that the number of operations needed to draw a sample with the required bias is in

$$\mathcal{O}\left(\frac{\sqrt{\|e^{-\tilde{B}_G^2 t}\| \alpha t^{3/2} D \log(f_{\max}/\epsilon_{TM})}}{\sqrt{\epsilon_{TM}} \gamma_M}\right). \quad (\text{G30})$$

Finally, if we set the error  $\epsilon_{TMH}$  to be the error also including the bias in the mean estimate of  $\dim \ker \Delta_k$  from having  $t = \log(1/\epsilon)/\gamma$ , we have, after choosing both

sources of error to be equal, that the systematic error can be made less than  $\epsilon_{TMH}$  using a number of operations in

$$\begin{aligned} & \tilde{\mathcal{O}}\left(\frac{\sqrt{\|e^{-\tilde{B}_G^2 t}\| \alpha D \log(f_{\max}/\epsilon_{TMH})}}{\sqrt{\epsilon_{TMH}} \gamma_M \gamma_{\min}^{3/2}}\right) \\ & \subseteq \tilde{\mathcal{O}}\left(\frac{D \sqrt{\alpha} \log(f_{\max}/\epsilon_{TMH})}{\sqrt{\epsilon_{TMH}} \gamma_M \gamma_{\min}^{3/2}}\right). \end{aligned} \quad (\text{G31})$$

### 3. Sample bounds

We finally need to consider the sampling error  $\epsilon_S$  that arises from taking only a finite number of samples. Standard probabilistic arguments show that the number of samples  $N_S$  needed to achieve a given  $\epsilon_S$  scales as  $\sigma^2/\epsilon_S^2$ . The mean-squared error  $\epsilon^2$  is then

$$\epsilon^2 = \epsilon_S^2 + \epsilon_{TMH}^2. \quad (\text{G32})$$

As before, we choose to make the two contributions to the error equal. There is a final source of complexity that needs to be considered though. Algorithm 1 begins by drawing a valid  $(k-1)$ -simplex to start at to ensure that we are within the space of interest. This means that we need to randomly draw vertices until we find a vertex that is in a  $k$ -clique. The probability of drawing such a simplex is  $|\text{Cl}_k(G)|/|\mathcal{H}_k|$ , which is the clique density for the graph. Thus with high probability, a number of samples proportional to the reciprocal of this will be needed. Each such sample requires clique detection, which is argued in Sec. III B scales as  $\mathcal{O}(|E|)$  up to logarithmic terms in  $k$ . This leads us to a cost of

$$N_{\text{op}} \in \tilde{\mathcal{O}}\left(\frac{|E| d_{k-1} \sigma^2}{|\text{Cl}_k(G)| \epsilon^2} + \frac{D \sqrt{\alpha} \log(f_{\max}/\epsilon)}{\epsilon^{5/2}} \frac{\sigma^2}{\gamma_M \gamma_{\min}^{3/2}}\right). \quad (\text{G33})$$

We now substitute  $\alpha, f_{\max}$  for variables related directly to the properties of the combinatorial Laplacian. We substitute  $t \geq \log(1/\epsilon)/\gamma_{\min}$  throughout and drop subdominant logarithmic terms. Firstly, from Eq. (G23) we get

$$\begin{aligned} \alpha & \leq 8rD\gamma_{\max}^3 \in \tilde{\Theta}\left(\frac{Dt^{3/2} \alpha^{1/2} \gamma_{\max}^3}{\epsilon^{1/2}}\right) \\ \implies \sqrt{\alpha} & \in \tilde{\Theta}\left(D \frac{\gamma_{\max}^3}{\gamma_{\min}^{3/2} \epsilon^{1/2}}\right). \end{aligned} \quad (\text{G34})$$

This bound on  $\sqrt{\alpha}$ , Eq. (G27), and the fact that  $\tilde{B}_G^2$  is positive semidefinite imply

$$r \in \tilde{\mathcal{O}}\left(\|e^{-\tilde{B}_G^2 t/2}\| \sqrt{\alpha} t^{3/2} / \sqrt{\epsilon}\right) = \tilde{\mathcal{O}}\left(D \gamma_{\max}^3 / \epsilon \gamma_{\min}^3\right). \quad (\text{G35})$$



From Eq. (G20), the logarithm scales as

$$\log\left(\frac{f_{\max}}{\epsilon}\right) \in \tilde{\Theta}\left(\log(d_{k-1}) + \frac{D^2 \gamma_{\max}^3}{\gamma_{\min}^3 \epsilon}\right). \quad (\text{G36})$$

The prior bound on the variance in Eq. (G10) evaluates to

$$\begin{aligned} \sigma^2 &\leq \frac{2^{2rD} (1/\epsilon)^{2D \max_i |\lambda_{p_i, \Gamma_i}| / \gamma_{\min}}}{d_{k-1}}, \\ &\in \frac{2^{\mathcal{O}(D^2 \kappa^3 / \epsilon)} (1/\epsilon)^{2D\kappa}}{d_{k-1}}, \end{aligned} \quad (\text{G37})$$

where we have defined the following quantity, which is analogous to the condition number for the Dirac operator restricted to  $\mathcal{H}_k$ ,  $\kappa = \gamma_{\max}/\gamma_{\min}$  (and neglecting the kernel). Substituting these expressions in Eq. (G33) then implies the number of operations for the algorithm obeys

$$\begin{aligned} N_{\text{op}} &\in \tilde{\mathcal{O}}\left(\frac{|E|d_{k-1}\sigma^2}{|\text{Cl}_k(G)|\epsilon^2} + \frac{D^4 \sigma^2 \gamma_{\max}^3}{\gamma_{\min}^3 \epsilon^3 \gamma_M} \left(\log(d_{k-1})D^{-2} + \frac{\gamma_{\max}^3}{\gamma^3 \epsilon}\right)\right), \\ &\in \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\epsilon^2} \left(\frac{|E|d_{k-1}}{|\text{Cl}_k(G)|} + \frac{D^4 \kappa^3}{\gamma_M \epsilon} \left(\log(d_{k-1})D^{-2} + \frac{\kappa^3}{\epsilon}\right)\right)\right). \end{aligned} \quad (\text{G38})$$

In the event we assume the worst-case bound on the variance, the total number of operations is in

$$N_{\text{op}} \in \tilde{\mathcal{O}}\left(\frac{2^{\mathcal{O}(D^2 \kappa^3 / \epsilon)}}{d_{k-1} \epsilon^{2+2D\kappa}} \left(\frac{|E|d_{k-1}}{|\text{Cl}_k(G)|} + \frac{D^4 \kappa^3}{\gamma_M \epsilon} \left(\log(d_{k-1})D^{-2} + \frac{\kappa^3}{\epsilon}\right)\right)\right). \quad (\text{G39})$$

This shows that under worst-case scenario scaling for the variance, our algorithm is efficient if  $D\kappa$  is polylogarithmic in  $n$ ,  $\kappa/\epsilon$  is a constant and  $\gamma_M^{-1}$  and the inverse density of cliques are polynomial in  $n$ .

While the above restrictions on the situations where the classical randomized algorithm is efficient are significant, they do imply that the TDA algorithm can be efficient even in cases where  $d_{k-1} = \binom{n}{k}$  is exponentially large provided the graph is clique dense. This possibility is not obvious if one compares only to classical algorithms like diagonalization, which scales polynomially with the dimension.

Finally, the number of operations varies in Eq. (G38) with the variance of the path integrals, which we upper bound with an exponential in  $D$ . While this scaling may seem prohibitive in the case where the graph is nearly complete, the variance bound in this case is extremely loose and using a particular bound designed for this scenario yields much better scaling as we will see later.

#### 4. Analysis

In order to give bounds on the sparsity of the combinatorial Laplacian, we first need to define a few terms. Let  $K$  be a simplicial complex with  $N$  vertices. The **up-degree**

of a  $k$ -simplex  $\sigma \in K$ , denoted by  $\deg_U(\sigma)$ , is the number of  $k+1$ -simplices in  $K$  that  $\sigma$  is in the boundary of. The **lower-degree** or **down-degree** of  $\sigma$ , denoted by  $\deg_L(\sigma)$ , is the number of  $k-1$ -simplices in  $K$  that are in the boundary of  $\sigma$ . If two  $k$ -simplices  $\sigma_1, \sigma_2 \in K$  both contain a  $k-1$ -simplex in their boundary, they are said to be **lower adjacent**. If  $\sigma_1, \sigma_2$  are in the boundary of a  $k+1$ -simplex, they are said to be **upper adjacent**. Lemma 3.2.4 of Ref. [55] shows that if  $\sigma_1, \sigma_2$  are distinct and lower adjacent, their common  $(k-1)$ -simplex is  $\sigma_1 \cap \sigma_2$  and is unique if it exists. Similarly, Lemma 3.2.2 of Ref. [55] shows that if  $\sigma_1$  and  $\sigma_2$  are upper adjacent, their common  $(k+1)$ -simplex is unique.

The down degree of a  $k$ -simplex  $\sigma \in K$  is always  $k+1$  for any simplicial complex simply because any  $k$ -simplex contains  $\binom{k+1}{k} = k+1$  simplices of dimension  $k-1$  in its boundary. Its up-degree is less trivial to determine, but can be bounded as follows.

*Proposition 10.*—Let  $K$  be a simplicial complex with  $N$  vertices and let  $\sigma$  be a  $k$ -simplex in  $K$ . The up-degree of  $\sigma$  is bounded by  $\min\{N-k-1, d\}$ , where  $d$  is maximum (up) degree of all the vertices in  $K$ .

*Proof.*—The most simple argument is that the up-degree is bounded by the number of possible ways to extend a  $k$ -simplex to a  $k+1$ -simplex by adding another point. Since

we have  $N - k - 1$  other points to choose from, the up-degree is bounded by this quantity.

A more rigorous argument is to consider the largest eigenvalue of the combinatorial Laplacian  $\Delta_k = \partial_k^\dagger \partial_k + \partial_{k+1} \partial_{k+1}^\dagger$  has the bound  $\lambda_{\max}(\Delta_k) \leq N$  [56]. The diagonal matrix elements of  $\Delta_k$  are given in Theorem 3.3.4 of Ref. [55] as  $\deg_U(\sigma_i) + k + 1$  for all the  $k$ -simplices  $\sigma_i \in K$  when  $k > 0$ . Since  $\Delta_k$  is a real symmetric matrix, the Courant-Fischer theorem [57] and the preceding bound on its largest eigenvalue together imply  $\deg_U(\sigma) + k + 1 \leq N$ , which in turn shows  $\deg_U(\sigma) \leq N - k - 1$ .

Another trivial upper bound is simply given by the maximal (up) degree  $d$  of any vertex  $v$  in the simplicial complex  $K$ . This is because the vertex added to a  $k$ -simplex to turn it into a  $(k + 1)$ -simplex has to be connected to all vertices in the  $k$ -simplex. The number of vertices connected to all the vertices in the  $k$ -simplex is upper bounded by  $d$ . Hence we may take  $\deg_U(\sigma) \leq \min\{N - k - 1, d\}$ . ■

*Proposition 11.*— $\partial_k$  has row-sparsity equal to  $\min\{N - k - 1, d\}$  and column sparsity equal to  $k + 1$ .  $\Delta_k$  has row and column sparsity bounded by  $(k + 1)(N - k - 1)$ .

*Proof.*—With the same notation as above,  $\partial_k$  acts on the vector space of  $k$ -simplices in  $K$  with the standard basis vectors corresponding to the  $k$ -simplices themselves. It has column sparsity equal to the down degree of any  $k$ -simplex in  $K$ , i.e.,  $k + 1$ , because  $\partial_k$  has nonzero elements in a column corresponding to a fixed  $k$ -simplex only when a  $k - 1$ -simplex is the boundary of that  $k$ -simplex.

Its row sparsity however is given by the largest up-degree of all  $k$ -simplices in  $K$  and from the preceding discussion is bounded by  $\min\{N - k - 1, d\}$ . As  $\Delta_k$  is Hermitian, the column and row sparsity of  $\Delta_k$  coincide and equals the maximum of the number of  $k$ -simplices  $c'$  such that  $c \cap c'$  is a  $(k - 1)$ -simplex and  $c \cup c'$  is *not* a  $(k + 1)$ -simplex for all  $k$ -simplices  $c$  (see Theorem 3.3.4 of Ref. [55]). A trivial upper bound for this is  $(k + 1)(N - k - 1)$  (note that this is precisely the product of the bounds given for the up and lower degrees earlier). This is because in order to construct for a given  $c$  a  $c'$  as above, we can remove any of the  $k + 1$  vertices from  $c$  and add any of the  $(N - k - 1)$  vertices not in  $c$  to form  $c'$ . ■

Note however that tighter bounds can be achieved by also using the largest degree  $d$  of a vertex, e.g., another upper bound on the sparsity is  $\mathcal{O}(kd)$ . Thus we can set  $D = \mathcal{O}(k^2 d^2)$  or  $D = \mathcal{O}(k^2(N - k)^2)$  for the number of terms  $D$  in the one-sparse decomposition of the combinatorial Laplacian depending on if  $d$  or  $N - k - 1$  is smaller [see the discussion preceding Eq. (G1)].

We now consider a few cases where our algorithm can run efficiently. Our first example will be the extreme case where the input to the algorithm is the completely disconnected graph on  $N$  points. The only nonzero Betti number of the associated clique complex in this case is  $\beta_0 = N$ . Since  $\dim \ker(\Delta_0) = \beta_0 = N$ ,  $\Delta_0$  is the  $N \times N$  zero matrix and  $\kappa$  is undefined. We therefore cannot directly

use the preceding asymptotic expressions for the number of samples required. The analysis is simple if we refer to the general formula for  $\sigma^2$  in Eq. (G10) however.

*Proposition 12.*—Let  $K$  be the completely disconnected graph on  $N$  vertices. The variance of the path-integral Monte Carlo sampling procedure for the Betti numbers of  $K$  is 0.

*Proof.*—In this case,  $\lambda = 0$  for all the eigenvalues of  $\Delta_0$ . We can choose the standard basis vectors for the  $N$ -dimensional vector space of 0-simplices, where each basis vector corresponds to a 0-simplex, as the eigenvectors of  $\Delta_0$ . Then  $W(\Gamma)$  is merely a product of Kronecker delta functions. There are precisely  $N$  valid paths (loops) for the Markov chain corresponding to each of the  $N$  vertices, so  $\sum_{\Gamma \in \mathcal{S}_\Gamma} W(\Gamma) = \sum_{\Gamma \in \mathcal{S}_\Gamma} |W(\Gamma)|^2 = \sum_{\Gamma \in \mathcal{S}_\Gamma} 1 = N$ . The variance then reduces to  $\sigma^2 = N^2/d_0^2 - N^2/d_0^2 = 0$ . ■

Even though the variance is precisely 0 for this case, we do not have *a priori* knowledge of the structure of the input graph from the standpoint of this algorithm, aside from an upper bound on the degree of the vertices. We possess only access to it via oracle queries that verify whether a collection of vertices forms a  $k$ -clique. Therefore, we cannot conclude anything definite about the Betti numbers with only a single sample in this situation as the zero variance result might suggest, unless given an additional promise that the *least* upper bound is a fixed value or 0.

Another simple case is when the input is a complete graph on  $N = n + 1$  points.

*Proposition 13.*—Let  $K$  be the clique complex of the complete graph on  $N = n + 1$  vertices. The variance of the path-integral Monte Carlo sampling procedure for the Betti numbers of  $K$  is 0.

*Proof.*—The clique complex of this graph corresponds to an  $n$ -simplex. Again from Theorem 3.3.4 of Ref. [55],  $(\Delta_k)_{ii} = \deg_U(\sigma_i) + k + 1$  and the off-diagonal entries are either 0 or  $\pm 1$  if  $k > 0$ . We argue that in this case, all the off-diagonal entries are 0. From the same theorem, it suffices to show that for a fixed  $k$ -simplex  $\sigma_i$ , the sum of the number of other upper adjacent simplices and other non-lower adjacent simplices must equal  $\binom{n+1}{k+1} - 1$  (since  $\Delta_k$  is a  $\binom{n+1}{k+1} \times \binom{n+1}{k+1}$  matrix).

We first calculate the up-degree of any  $k$ -simplex in our  $n$ -simplex. The number of  $(k + 1)$ -simplices in an  $n$ -simplex is  $\binom{n+1}{k+2}$ . The number of  $k$ -simplices in a  $(k + 1)$ -simplex is  $k + 2$ . The number of  $k$ -simplices in an  $n$ -simplex is  $\binom{n+1}{k+1}$ . Intuitively, we can determine the up-degree by finding the number of  $k$ -simplices contained in the boundary of all  $(k + 1)$ -simplices and dividing by the actual number of  $k$ -simplices in an  $n$ -simplex to account for overcounting. This quantity is precisely

$$\frac{(k + 2)\binom{n+1}{k+2}}{\binom{n+1}{k+1}} = n - k$$

and we thus get that the diagonal entries of  $\Delta_k$  are all equal to  $n + 1$  when  $k > 0$  (note that this saturates the bound on the up-degree of a  $k$ -simplex in any simplicial complex given earlier with  $N = n + 1$ ).

Now fix a particular  $k$ -simplex  $\sigma$  in the  $n$ -simplex. We want to count the number of *other*  $k$ -simplices that are *not* lower adjacent to  $\sigma$ .  $\sigma$  has  $k + 1$ -simplices of dimension  $k - 1$  in its boundary. From the above result, each of these  $(k - 1)$ -simplices has up-degree  $n - (k - 1) = n - k + 1$ . Then the number of potential lower adjacent  $k$ -simplices is  $(k + 1)(n - k + 1)$ . But by the uniqueness of common upper simplices, we have counted  $\sigma$   $k + 1$  times, one for each of the  $(k - 1)$ -simplices in the boundary of  $\sigma$ . Thus the number of *other* lower adjacent simplices is  $(k + 1)(n - k + 1) - (k + 1) = (k + 1)(n - k)$ . Thus the number of *other*  $k$ -simplices *not* lower adjacent to  $\sigma$  is  $\binom{n+1}{k+1} - (k + 1)(n - k) - 1$ . On the other hand, the number of *other*  $k$ -simplices that are upper adjacent to  $\sigma$  is given by multiplying the up-degree  $n - k$  of  $\sigma$ , which gives the number of  $k + 1$ -simplices that are upper adjacent to  $\sigma$ , by the number of  $k$ -simplices in each  $k + 1$ -simplex, which is  $k + 2$ . But this overcounts  $\sigma$  by  $n - k$ , so  $(n - k)(k + 2) - (n - k) = (n - k)(k + 1)$ . The sum of this and  $\binom{n+1}{k+1} - (k + 1)(n - k) - 1$  is clearly  $\binom{n+1}{k+1} - 1$ , so the only nonzero entries in  $\Delta_k$  are the diagonal ones when  $k > 0$ .

When  $k = 0$ , Theorem 3.3.4 in Ref. [55] shows the diagonal entries of  $\Delta_0$  are the degree of the vertices in our  $n$ -simplex, which is precisely  $n$ . It also implies all the off-diagonal entries are 0 since every vertex is upper adjacent in an  $n$ -simplex. Thus in either the  $k = 0$  or  $k > 0$  case,  $\Delta_k$  is proportional to the identity. We can then set  $D = r = 1$  in Eq. (G10) and the number of valid paths summed over is precisely  $d_k$ . By an analogous reasoning to the completely disconnected case,  $\sum_{\Gamma \in \mathcal{S}_\Gamma} W(\Gamma) = \sum_{\Gamma \in \mathcal{S}_\Gamma} |W(\Gamma)|^2 = \sum_{\Gamma \in \mathcal{S}_\Gamma} d_k$  and

$$\sigma^2 = e^{-2i\lambda} \left( \frac{d_k^2}{d_k^2} - \frac{d_k^2}{d_k^2} \right) = 0,$$

where  $\lambda = n + 1$  when  $k > 0$  and  $\lambda = n$  when  $k = 0$ . ■

Analogous arguments on needing a promise on the least upper bound on the degree of the vertices hold as in the case of the completely disconnected graph above.

More can be said of  $\kappa$  in restricted cases. For instance, it is known that orientable simplicial complexes of dimension  $d \leq 2$  have  $\kappa \in \mathcal{O}(n_k^2)$ , where  $n_k$  denotes the number of  $k$ -simplices and  $0 \leq k \leq 2$ , and it is further conjectured that  $\kappa \in \mathcal{O}(n_k^{2/d})$  in most cases [14]. This latter bound is quite favorable for high-dimensional simplices, i.e., when  $d \sim n$  and with the approximation that  $n_k \sim 2^n$ . Unfortunately, it is difficult to give tight bounds on  $\gamma_M$  and  $\sigma^2$  in Eq. (G38) for arbitrary simplicial complexes. Though the Perron-Frobenius theorem applied to stochastic matrices

implies  $0 < \gamma_M \leq 1$  [57], its scaling with the other parameters of relevance are unknown.

While we do not ultimately expect this classical TDA method to be efficient generically, the above discussions show there exist sufficient conditions under which it can be used to extract Betti numbers in polynomial time without suffering from a generic exponential dependence on the number of data points characteristic of other classical TDA algorithms. Furthermore, it creates uncertainty about the necessary and sufficient conditions for an exponential advantage for quantum TDA as the cases where this algorithm has an exponential advantage relative to the deterministic classical algorithm are analogous to those for the quantum algorithm. This means that an identification of clear cases where exponential advantage is likely remains an important open problem within the domain of quantum algorithms for TDA.

- 
- [1] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd, Quantum machine learning, *Nature* **549**, 195 (2017).
  - [2] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost, Quantum principal component analysis, *Nat. Phys.* **10**, 631 (2014).
  - [3] Aram W. Harrow, Avinatan Hassidim, and Seth Lloyd, Quantum algorithm for linear systems of equations, *Phys. Rev. Lett.* **103**, 150502 (2009).
  - [4] Ewin Tang, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, New York, 2019), p. 217.
  - [5] Ewin Tang, Quantum principal component analysis only achieves an exponential speedup because of its state preparation assumptions, *Phys. Rev. Lett.* **127**, 060503 (2021).
  - [6] Yuval R. Sanders, Dominic W. Berry, Pedro C. S. Costa, Louis W. Tessler, Nathan Wiebe, Craig Gidney, Hartmut Neven, and Ryan Babbush, Compilation of fault-tolerant quantum heuristics for combinatorial optimization, *PRX Quantum* **1**, 020312 (2020).
  - [7] Ryan Babbush, Jarrod R. McClean, Michael Newman, Craig Gidney, Sergio Boixo, and Hartmut Neven, Focus beyond quadratic speedups for error-corrected quantum advantage, *PRX Quantum* **2**, 010103 (2021).
  - [8] Seth Lloyd, Silvano Garnerone, and Paolo Zanardi, Quantum algorithms for topological and geometric analysis of data, *Nat. Commun.* **7**, 10138 (2016).
  - [9] Sam Gunn and Niels Kornerup, Review of a quantum algorithm for Betti numbers, [arXiv:1906.07673](https://arxiv.org/abs/1906.07673) (2019).
  - [10] Chris Cade and P. Marcos Crichigno, Complexity of supersymmetric systems and the cohomology problem, [arXiv:2107.00011](https://arxiv.org/abs/2107.00011) (2021).
  - [11] Shashanka Ubaru, Ismail Yunus Akhalwaya, Mark S. Squillante, Kenneth L. Clarkson, and Lior Horesh, Quantum topological data analysis with linear depth and exponential speedup, [arXiv:2108.02811](https://arxiv.org/abs/2108.02811) (2021).
  - [12] Ryu Hayakawa, Quantum algorithm for persistent Betti numbers and topological data analysis, *Quantum* **6**, 873 (2022).

- [13] Beno Eckmann, Harmonische funktionen und randwertaufgaben in einem komplex, *Commentarii Mathematici Helvetici* **17**, 240 (1944/45).
- [14] Joel Friedman, Computing Betti numbers via combinatorial Laplacians, *Algorithmica* **21**, 331 (1998).
- [15] Casper Gyurik, Chris Cade, and Vedran Dunjko, Towards quantum advantage via topological data analysis, *Quantum* **6**, 855 (2022).
- [16] Nai-Hui Chia, András Gilyén, Tongyang Li, Han-Hsuan Lin, Ewin Tang, and Chunhao Wang, in *Proceedings of the 52nd Annual ACM SIGACT symposium on theory of computing* (Association for Computing Machinery, New York, 2020), p. 387.
- [17] Marcos Crichigno and Tamara Kohler, Clique homology is QMA1-hard, [arXiv:2209.11793](https://arxiv.org/abs/2209.11793) (2022).
- [18] Michał Adamaszek and Juraj Stacho, Complexity of simplicial homology and independence complexes of chordal graphs, *Comput. Geom.* **57**, 8 (2016).
- [19] Peter Scheiblechner, On the complexity of deciding connectedness and computing Betti numbers of a complex algebraic variety, *J. Complex.* **23**, 359 (2007).
- [20] Bojan Mohar, Isoperimetric numbers of graphs, *J. Comb. Theory, Ser. B* **47**, 274 (1989).
- [21] Anna Gundert and May Szedlák, Higher dimensional discrete Cheeger inequalities, *J. Comput. Geometry* **6**, 54 (2015).
- [22] Theodore J. Yoder, Guang Hao Low, and Isaac L. Chuang, Fixed-point quantum search with an optimal number of queries, *Phys. Rev. Lett.* **113**, 210501 (2014).
- [23] Guang Hao Low and Isaac L. Chuang, Hamiltonian simulation by qubitization, *Quantum* **3**, 163 (2019).
- [24] Dominic W. Berry, Mária Kieferová, Artur Scherer, Yuval R. Sanders, Guang Hao Low, Nathan Wiebe, Craig Gidney, and Ryan Babbush, Improved techniques for preparing eigenstates of fermionic Hamiltonians, *Npj Quantum Inf.* **4**, 22 (2018).
- [25] Andreas Bärttschi and Stephan Eidenbenz, in *Fundamentals of Computation Theory*, edited by Leszek Antoni Gasieniec, Jesper Jansson, and Christos Levkopoulos (Springer International Publishing, Cham, 2019), p. 126.
- [26] Sara Ayman Metwalli, François Le Gall, and Rodney Van Meter, Finding small and large  $k$ -clique instances on a quantum computer, *IEEE Trans. Quantum Eng.* **1**, 3102911 (2020).
- [27] Ian D. Kivlichan, Craig Gidney, Dominic W. Berry, Nathan Wiebe, Jarrod McClean, Wei Sun, Zhang Jiang, Nicholas Rubin, Austin Fowler, Alán Aspuru-Guzik, Hartmut Neven, and Ryan Babbush, Improved fault-tolerant quantum simulation of condensed-phase correlated electrons via trotterization, *Quantum* **4**, 296 (2020).
- [28] Gilles Brassard, Peter Høyer, and Alain Tapp, in *Automata, Languages and Programming* (Springer, Berlin, Heidelberg, 1998), p. 820.
- [29] Michel Boyer, Gilles Brassard, Peter Høyer, and Alain Tapp, Tight bounds on quantum searching, *Fortschr. Phys.* **46**, 493 (1998).
- [30] Ryan Babbush, Craig Gidney, Dominic W. Berry, Nathan Wiebe, Jarrod McClean, Alexandru Paler, Austin Fowler, and Hartmut Neven, Encoding electronic spectra in quantum circuits with linear T complexity, *Phys. Rev. X* **8**, 041015 (2018).
- [31] Aram W. Harrow and Richard A. Low, Random quantum circuits are approximate 2-designs, *Commun. Math. Phys.* **291**, 257 (2009).
- [32] Dominic W. Berry, Andrew M. Childs, and Robin Kothari, in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on* (IEEE, Berkeley, 2015), p. 792.
- [33] Guang Hao Low and Isaac L. Chuang, Optimal Hamiltonian simulation by quantum signal processing, *Phys. Rev. Lett.* **118**, 010501 (2017).
- [34] Ryan Babbush, Dominic W. Berry, and Hartmut Neven, Quantum simulation of the Sachdev-Ye-Kitaev model by asymmetric qubitization, *Phys. Rev. A* **99**, 040301 (2019).
- [35] Shashanka Ubaru and Yousef Saad, in *International Conference on Machine Learning* (PMLR, New York, 2016), p. 468.
- [36] Simon Apers, Sander Gribling, Sayantan Sen, and Dániel Szabó, A (simple) classical algorithm for estimating Betti numbers, *Quantum* **7**, 1202 (2023).
- [37] Michał Adamaszek, Extremal problems related to Betti numbers of flag complexes, *Discrete Appl. Math.* **173**, 8 (2014).
- [38] Joonho Lee, Dominic W. Berry, Craig Gidney, William J. Huggins, Jarrod R. McClean, Nathan Wiebe, and Ryan Babbush, Even more efficient quantum computations of chemistry through tensor hypercontraction, *PRX Quantum* **2**, 030305 (2021).
- [39] Matthew Kahle, Topology of random clique complexes, *Discrete Math.* **309**, 1658 (2009).
- [40] Michael Goff, Extremal Betti numbers of Rips complexes, [arXiv:0910.0040](https://arxiv.org/abs/0910.0040) (2009).
- [41] Matthew Kahle, Random geometric complexes, *Discrete Comput. Geom.* **45**, 553 (2011).
- [42] David Cohen-Steiner, Weihao Kong, Christian Sohler, and Gregory Valiant, in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, New York, 2018), p. 1263.
- [43] Sayan Mukherjee and John Steenbergen, Random walks on simplicial complexes and harmonics, *Random Struct. Algor.* **49**, 379 (2016).
- [44] Ori Parzanchevski and Ron Rosenthal, Simplicial complexes: Spectrum, homology, and random walks, *Random Struct. Algor.* **50**, 225 (2017).
- [45] Dominic W. Berry, Graeme Ahokas, Richard Cleve, and Barry C. Sanders, Efficient quantum algorithms for simulating sparse Hamiltonians, *Commun. Math. Phys.* **270**, 359 (2007).
- [46] Dominic W. Berry, Andrew M. Childs, Richard Cleve, Robin Kothari, and Rolando D. Somma, in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing, STOC '14* (ACM, New York, NY, USA, 2014), p. 283.
- [47] John Lee, *Introduction to Topological Manifolds* (Springer Science & Business Media, New York, 2010), Vol. 202.
- [48] Melissa R. McGuirl, Homology and persistent homology bootcamp notes summer @ ICERM 2017, (2017).
- [49] Robert Ghrist, Barcodes: The persistent topology of data, *Bull. Am. Math. Soc.* **45**, 61 (2008).

- [50] Afra Zomorodian and Gunnar Carlsson, Computing persistent homology, [Discrete Comput. Geom.](#) **33**, 249 (2005).
- [51] Herbert Edelsbrunner and John L. Harer, *Computational Topology: an Introduction* (American Mathematical Society, Providence, Rhode Island, 2022).
- [52] Hiroshi Imai and Masahito Hayashi, Fourier analytic approach to phase estimation in quantum systems, [New J. Phys.](#) **11**, 043034 (2009).
- [53] Olle Häggström, *Finite Markov Chains and Algorithmic Applications* (Cambridge University Press, Cambridge, UK, 2002).
- [54] Andrew M. Childs, Yuan Su, Minh C. Tran, Nathan Wiebe, and Shuchen Zhu, Theory of Trotter error with commutator scaling, [Phys. Rev. X](#) **11**, 011020 (2021).
- [55] Timothy E. Goldberg, Combinatorial Laplacians of simplicial complexes, Senior Thesis, Bard College (2002).
- [56] Art Duval and Victor Reiner, Shifted simplicial complexes are Laplacian integral, [Trans. Am. Math. Soc.](#) **354**, 4313 (2002).
- [57] Roger A. Horn and Charles R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, UK, 2012), 2nd ed.