nature COMMUNICATIONS

ARTICLE

Received 17 Sep 2014 | Accepted 9 Nov 2015 | Published 25 Jan 2016

DOI: 10.1038/ncomms10138 OPEN

Quantum algorithms for topological and geometric analysis of data

Seth Lloyd[1], Silvano Garnerone[2] & Paolo Zanardi[3]

Extracting useful information from large data sets can be a daunting task. Topological methods for analysing data sets provide a powerful technique for extracting such information. Persistent homology is a sophisticated tool for identifying topological features and for determining how such features persist as the data is viewed at different scales. Here we present quantum machine learning algorithms for calculating Betti numbers—the numbers of connected components, holes and voids—in persistent homology, and for finding eigenvectors and eigenvalues of the combinatorial Laplacian. The algorithms provide an exponential speed-up over the best currently known classical algorithms for topological data analysis.

[1] Department of Mechanical Engineering, Research Lab for Electronics, Massachusetts Institute of Technology, MIT 3-160, Cambridge, Massachusetts 02139, USA. [2] Institute for Quantum Computing, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. [3] Department of Physics and Astronomy, Center for Quantum Information Science & Technology, University of Southern California, Los Angeles, California 90089-0484, USA. Correspondence and requests for materials should be addressed to S.L. (email: slloyd@mit.edu).

Human society is currently generating on the order of Avogadro's number (6×10^{23}) of bits of data a year. Extracting useful information from even a small subset of such a huge data set is difficult. A wide variety of big data processing techniques have been developed to extract from large data sets the hidden information in which one is actually interested. Topological techniques for analysing big data represent a sophisticated and powerful tool[1–24]. By its very nature, topology reveals features of the data that robust to how the data were sampled, how it was represented and how it was corrupted by noise. Persistent homology is a particularly useful topological technique that analyses the data to extract topological features such as the number of connected components, holes, voids and so on (Betti numbers) of the underlying structure from which the data was generated. The length scale of analysis is then varied to see whether those topological features persist at different scales. A topological feature that persists over many length scales can be identified with a 'true' feature of the underlying structure.

Topological methods for analysis face challenges: a data consisting of n data points possesses 2^n possible subsets that could contribute to the topology. Performing methods of algebraic topology on simplicial complexes eventually requires matrix multiplication or diagonalization of matrices of dimension $O\binom{n}{k+1}$ to extract topological features at dimension k. For small k, such operations require time polynomial in n; however, to extract high-dimensional features, matrix multiplication and diagonalization lead to problem solution scalings that grow exponentially in the size of the complex. A variety of mathematical methods have been developed to cope with the resulting combinatorial explosion, notably mapping the complex to a smaller complex with the same homology, and then performing the matrix operations on the reduced complex[1–24]. Even in such cases, the initial reduction must identify all simplices in the original complex, and so can scale no better than linearly in the number of simplices. Consequently, even with only a few hundred data points, creating the persistent homology for Betti numbers at all orders of k is a difficult task. In particular, the most efficient classical algorithms for estimating Betti numbers at order k (the number of k-dimensional gaps, holes and so on), have computational complexity either exponential in k or exponential in n (refs 7–12), so that estimating Betti numbers to all orders scales exponentially in n, and algorithms for diagonalizing the combinatorial Laplacian (that reveal not only the Betti numbers but additional geometric structure) at order k have computational complexity as $O\left(\binom{n}{k}^2\right)$, where n is the number of vertices in the (possibly reduced) complex. That is, the best classical algorithms for estimating Betti numbers to all orders[9–12] and for diagonalizing the full combinatorial Laplacian grow exponentially in the number of vertices in the complex.

This paper investigates quantum algorithms for performing topological analysis of large data sets. We show that a quantum computer can find the eigenvectors and eigenvalues of the combinatorial Laplacian and estimate Betti numbers to all orders and to accuracy δ in time $O(n^5/\delta)$, thereby reducing a classical problem for which the best existing solutions have exponential computational complexity, to a polynomial-time quantum problem. Betti numbers can also be estimated by using a reduced, or 'witness' complex, that contains fewer points than the original complex[1–12]. Applied to such witness complexes, our method again yields a reduction in estimation time from $O(2^{2\tilde{n}})$ to $O(\tilde{n}^5)$, where $\tilde{n}$ is the number of points in the reduced complex.

Recently, quantum mechanical techniques have been proposed for machine learning and data analysis[25–34]. In particular, some quantum machine learning algorithms[31–33] provide exponential speed-ups over the best existing classical algorithms for supervised and unsupervised learning. Such 'big quantum data' algorithms use a quantum random access memory (qRAM)[35–37] to map an N-bit classical data set onto the quantum amplitudes of a $(\log_2 N)$-qubit quantum state, an exponential compression over the classical representation. The resulting state is then manipulated using quantum information processing in time $\mathrm{poly}(\log_2 N)$ to reveal underlying features of the data set. That is, quantum computers that can perform 'quantum sampling' of data can perform certain machine learning tasks exponentially faster than classical computers performing classical sampling of data. A discussion of computational complexity in quantum machine learning can be found in ref. 34. Constructing a large-scale qRAM to access $N\sim10^9-10^{12}$ pieces of data is a difficult task. By contrast, the topological and geometrical algorithms presented here do not require a large-scale qRAM: a qRAM with $O(n^2)$ bits suffices to store all pairwise distance information between the points of our data set. The algorithms presented here obtain their exponential speed-up over the best existing classical algorithms not by having quantum access to a large data set, but instead, by mapping a combinatorially large simplicial complex with $O(2^n)$ simplices to a quantum state with n qubits, and by using quantum information processing techniques such as matrix inversion and diagonalization to perform topological and geometrical analysis exponentially faster than classical algorithms. Essentially, our quantum algorithms operate by finding the eigenvectors and eigenvalues of the combinatorial Laplacian. But diagonalizing a 2^n by 2^n sparse matrix using a quantum computer takes time $O(n^2)$, compared with time $O(2^{2n})$ on a classical computer[38–40].

The algorithms given here are related to quantum matrix inversion algorithms[41]. The original matrix inversion algorithm[41] yielded as solution a quantum state, and left open the question of how to extract useful information from that state. The topological and geometric algorithms presented here answer that question: the algorithms yield as output not quantum states but rather topological invariants—Betti numbers—and do so in time exponentially faster than the best existing classical algorithms. The best classical algorithms for calculating the kth Betti number takes time $O(n^k)$, and estimating Betti numbers to all orders to accuracty δ takes time at least $O(2^n\log(1/\delta))$ (refs 7–12). Exact calculation of Betti numbers for some types of topological sets (algebraic varieties) is PSPACE hard[42]. By contrast, our algorithm provides approximate values of Betti numbers to all orders and to accuracy δ in time $O(n^5/\delta)$: although no polynomial classical algorithm for such approximate evaluation of topological invariants is known, the computational complexity of such approximation remains an open problem. We do not expect our quantum algorithms to solve a PSPACE-hard problem in polynomial time. We summarize the comparison between the amount of resources required by the classical and quantum algorithms in Table 1.

Results

The quantum pipeline. The quantum algorithm operates by mapping vectors, simplices, simplicial complexes and collections of simplicial complexes to quantum mechanical states, and reveals topology by performing linear operations on those states. The 2^n possible simplices of the simplicial complex are mapped onto an n-qubit quantum state. This state is then analysed using conventional quantum computational techniques of eigenvector and eigenvalue analysis, matrix inversion and so on. The quantum analysis reveals topological features of the data, and shows how those features arise and persist when the scale of analysis is varied. The resulting quantum algorithms provide an

exponential speed-up over the best existing classical algorithms for topological data analysis.

In addition to revealing topological features such as Betti numbers, our algorithm uses the relationship between algebraic topology and Hodge theory[9–12,14–24] to reveal geometrical information about the data analysed at different scales. The algorithm operates by identifying the harmonic forms of the data, together with the other eigenvalues and eigenvectors of the combinatorial Laplacian—the quantities that famously allow one to 'hear the shape of a drum'[43]. The quantum algorithm reveals these geometric features exponentially faster than the corresponding classical algorithms. In particular, our quantum algorithm for finding all Betti numbers for the persistent homology for simplicial complexes over n points and for diagonalizing the combinatorial Laplacian takes time $O(n^5/\delta)$, where δ is the multiplicative accuracy to which Betti numbers and eigenvalues are determined. The best available classical algorithms to perform these tasks at all orders of k take time $O(2^{2n}\log(1/\delta))$.

The advantage of big quantum data techniques is that they provide exponential compression of the representation of the data. The challenge is to see if—and this is a big 'if'—it is still possible to process the highly compressed quantum data to reveal the desired hidden structure that underlies the original data set. Here we show that quantum information processing acting on large data sets encoded in a quantum form can indeed reveal topological features of the data set.

Classical algorithms for persistent homology have two steps (the 'pipeline'). First, one processes the data to allow the construction of a topological structure such as a simplicial complex that approximates the hidden structure from which the data was generated. The details of the topological structure depends on the scale at which data is grouped together. Second, one constructs topological invariants of that structure and analyses how those invariants behave as a function of the grouping scale. As above, topological invariants that persist over a wide range of scales are identified as features of the underlying hidden structure.

The quantum 'pipeline' for persistent homology also has two steps. First, one accesses the data in quantum parallel to construct quantum states that encode the desired topological structure: if the structure is a simplicial complex, for example, one constructs quantum states that are uniform superposition of descriptions of the simplices in the complex. Second, one uses the ability of quantum computing to reveal the ranks of linear maps to construct the topological invariants of the structure. The steps of the quantum pipeline are now described in more detail.

Constructing a simplicial complex. Classical persistent homology algorithms use the access to data and distances to construct a topological structure—typically a simplicial complex—that corresponds to the hidden structure whose topology one wishes to reveal. In the quantum algorithm, we use the ability to access data and to estimate distances in quantum parallel to construct quantum states that encode the simplicial complex. Each simplex in the complex consists of a fully connected set of vertices: a k-simplex s_k consists of $k+1$ vertices $j_0, j_1, \ldots, j_k$ (listed in ascending order, $j_0<j_1< \ldots <j_k$) together with the $k(k+1)/2$ edges connecting each vertex to all the other vertices in the simplex. Encode a k-simplex s_k as a string of n bits, for example, 0110 … 1, with $k+1$ 1s at locations $j_0, j_1, \ldots, j_k$ designating the vertices in the simplex. Removing the ℓth vertex and its associated edges from a k-simplex yields a $k-1$ simplex. The $k+1$ simplices $s_{k-1}(\ell)$ with vertices $j_0 \ldots \hat{j}_\ell \ldots j_k$ obtained by removing the ℓth vertex j_ℓ from s_k form the boundary of the original simplex. The number of potential simplices in a simplicial complex is equal to 2^n, the number of possible subsets of the n points in the graph. That is, every member of the power set is a potential simplex. If n is large, the resulting combinatorial explosion means that identifying large simplices can be difficult.

To define a simplicial complex, fix a grouping scale ϵ, and identify k simplices as subsets of $k+1$ points that are all within ϵ of each other. The resulting set of simplices S^ϵ is called the Vietoris–Rips complex. The form of the simplicial complex S^ϵ depends on the scale ϵ at which its points are grouped together: persistent homology investigates how topological invariants of the simplicial complex depend on the scale ϵ. The collection of simplicial complexes $\{S^\epsilon\}$ for different values of the grouping scale ϵ is called a filtration. Note that if a simplex belongs to the complex S^ϵ, then it also belongs to $S^{\epsilon'}$, $\epsilon'>\epsilon$. That is, the filtration consists of a sequence of nested simplicial complexes. When ϵ is sufficiently small, only the zero-simplices (points) lie in the complex. As ϵ increases, one and two simplices (edges and triangles) enter the complex, followed by higher order simplices. As ϵ continues to increase, topological features such as holes, gaps and voids come into existence, and then are eventually filled in. For sufficiently large ϵ, all possible simplices are contained in the complex.

Now construct quantum states that correspond to the simplicial complex. Encode simplices as quantum states over n qubits with 1 s at the positions of the vertices. We designate the k-simplex s_k by the n-qubit basis vector $|s_k\rangle \in C^{2^n}$. Denote the $\binom{n}{k+1}$ dimensional Hilbert space corresponding to all possible k simplices by W_k. Let $\mathcal{H}_k^\epsilon$ be the subspace of W_k spanned by $|s_k\rangle$ where $s_k \in S_k^\epsilon$, the set of k simplices in S^ϵ. The full simplex space at scale ϵ is defined to be $\mathcal{H}^\epsilon = \oplus_k H_k^\epsilon$. Assume that the distances between pairs of points are either given by a quantum algorithm or stored in qRAM (see Methods section). The ability to evaluate distances translates onto the ability to apply the projector P_k^ϵ that projects onto the k-simplex space $\mathcal{H}_k^\epsilon$and the projector P^ϵ that projects onto the full simplex space $\mathcal{H}^\epsilon$.

Table 1 | Computational cost comparison.

Procedural steps	Classical cost	Quantum cost
Input pairwise distances, n points	$O(n^2)$ bits	$O(n^2)$ bits
Construct simplicial complex	$O(2^n)$ ops	$O(n^2)$ ops on $O(n)$ qubits
Diagonalize Laplacian/find Betti numbers	$O(2^{2n}\log(1/\delta))$ ops	$O(n^5/\delta)$ quantum ops

δ is the multiplicative accuracy to which the Betti numbers and the eigenvalues of the combinatorial Laplacian are determined. Note the trade-off between the exponential quantum speed-up and accuracy: the quantum algorithms obtain an exponential speed-up over classical algorithms but provide an accuracy that scales polynomially in $1/\delta$ rather than exponentially. This feature arises from the nature of the quantum phase estimation/matrix inversion algorithms, which obtain their exponential speed-up by estimating eigenvectors and eigenvalues using a 'pointer-variable' measurement interaction[38–40]. By contrast, classical algorithms need only keep $O(\log(1/\delta))$ bits of precision, but must perform $O(2^{2n})$ steps to diagonalize $2^n \times 2^n$ sparse matrices.

Grover's algorithm can then be used to construct the k-simplex state

$$|\psi\rangle_k^\epsilon = \frac{1}{\sqrt{|S_k^\epsilon|}} \sum_{s_k \in S_k^\epsilon} |s_k\rangle, \quad (1)$$

where as above S_k^ϵ is the set of k simplices in the complex at scale ϵ. That is, $|\psi\rangle_k^\epsilon$ is the uniform superposition of the quantum states corresponding to k simplices in the complex. For each simplex s_k, we can verify whether $s_k \in S_k^\epsilon$ in $O(k^2)$ steps. That is, we can implement a membership function $f_k^\epsilon(s_k) = 1$ of $s_k \in S_k^\epsilon$ in $O(k^2)$ steps. The multi-solution version of Grover's algorithm then allows us to construct the k-simplex state of equation (1).

The construction of the k-simplex state via Grover's algorithm reveals the number of k simplices $|S_k^\epsilon| = \dim H_k^\epsilon$ in the complex at scale ϵ, and takes time $O\left(n^2 (\zeta_k^\epsilon)^{-1/2}\right)$, where $\zeta_k^\epsilon = |S_k^\epsilon| / \binom{n}{k+1}$ is the fraction of possible k simplices that are actually n the complex at scale ϵ. When this fraction is too small, the quantum search procedure will fail to find the simplices. For $k \ll n$, we have $\binom{n}{k+1} = O(n^{k+1}/k!)$, and ζ_k^ϵ is only polynomially small in n. By contrast, for $k \approx n$, ζ_k^ϵ can be exponentially small in n: if only an exponentially small set of possible simplices actually lie in the complex, quantum search will fail to find them. For the purposes of performing the quantum algorithm, we fix a parameter ζ that determines the accuracy to which we wish to determine the simplex state, and run the simplex finding algorithm for a time $\zeta^{-1/2}$. At each grouping scale ϵ, the algorithm will find k simplices when $\zeta_k^\epsilon > \zeta$, and estimate the number of k simplices to accuracy $\zeta_k^\epsilon \pm \zeta$. As ϵ increases, more and more simplices enter into the complex; ζ_k^ϵ increases; and quantum search will succeed in constructing the simplex state to greater and greater accuracy. When ϵ becomes larger than the maximum distance between vectors, all simplices are in the complex.

Below, it will prove useful to have, in addition to the simplex state $|\psi\rangle_k^\epsilon$ the state $\rho_k^\epsilon = \left(1/|S_k^\epsilon|\right) \sum_{s_k \in S_k^\epsilon} |s_k\rangle\langle s_k|$, which is the uniform mixture of all k-simplex states in the complex at grouping scale ϵ. ρ_k^ϵ can be constructed in a straightforward fashion from the simplex state $|\psi\rangle_k^\epsilon$ by adding an ancilla and copying the simplex label to construct the state $\frac{1}{\sqrt{|S_k^\epsilon|}} \sum_{s_k \in S_k^\epsilon} |s_k\rangle \otimes |s_k\rangle$. Tracing out the ancilla then yields the desired uniform mixture over all k simplices.

In summary, we can represent the the simplicial complex in quantum mechanical form using exponentially fewer bits than that are required classically. Indeed, the quantum search method for constructing simplicial states works best when ζ_k^ϵ is not too small, so that a substantial fraction of simplices that could be in the complex are actually in the complex. But this regime is exactly the regime where the classical algorithms require an exponentially large amount of memory space bits merely to record which simplices are in the complex. Now we show how to act on this quantum mechanical representation of the filtration to reveal persistent homology.

Topological analysis. Having constructed a quantum state that represents the simplicial complex S^ϵ at scale ϵ, we use quantum information processing to analyse its topological properties. In algebraic topology in general, and in persistent homology in particular, this analysis is performed by investigating the properties of linear maps on the space of simplices. As above, let $\mathcal{H}_k^\epsilon$ be the Hilbert space spanned by vectors corresponding to k simplices in the complex at level ϵ. We identify the vector space $\mathcal{H}_k^\epsilon$ with the abelian group C_k (the kth chain group) under addition of vectors in the space. Let $j_0 \ldots j_k$ be the vertices of s_k. Define the boundary map ∂_k on the space of k simplices by

$$\partial_k |s_k\rangle = \sum_\ell (-1)^\ell |s_{k-1}(\ell)\rangle \quad (2)$$

where as above $s_{k-1}(\ell)$ is the $k-1$ simplex on the boundary of s_k with vertices $j_0 \ldots \hat{j}_\ell \ldots j_k$ obtained by omitting the ℓth vertex j_ℓ from s_k. The boundary map maps each simplex to the oriented sum of its boundary simplices. ∂_k is a $\binom{n}{k} \times \binom{n}{k+1}$ matrix with $n-k$ non-zero entries ± 1 in each row and $k+1$ non-zero entries ± 1 per column. Note that $\partial_k \partial_{k+1} = 0$: the boundary of a boundary is zero. As defined, ∂_k acts on the space of all k simplices. We define the boundary map restricted to operate from $\mathcal{H}_k^\epsilon$ to H_{k-1}^ϵ to be $\tilde{\partial}_k = \partial_k P_k^\epsilon$, where as above P_k^ϵ is the projector onto the space of k simplices in the complex at scale ϵ.

The kth homology group $\mathbf{H}_k$ is the quotient group, $\mathrm{Ker}\,\tilde{\partial}_k / \mathrm{Image}_{k+1}\tilde{\partial}_{k+1}$, the kernel of $\tilde{\partial}_k$ divided by the image of $\tilde{\partial}_{k+1}$ acting on $\mathcal{H}_{k+1}$at grouping scale ϵ. The kth Betti number β_k is equal to the dimension of $\mathbf{H}_k$, which in turn is equal to the dimension of the kernel of $\tilde{\partial}_k$ minus the dimension of the image of $\tilde{\partial}_{k+1}$.

The strategy that we use to identify persistent topological features operates by identifying the singular values and singular vectors of the boundary map. Connected components, holes, voids and so on, correspond to structures—chains of simplices—that have no boundary, but that are not themselves a boundary. That is, we are looking for the set of states that lie within the kernel of $\tilde{\partial}_k$, but that do not lie within the image of $\tilde{\partial}_{k+1}$. The ability to decompose arbitrary vectors in $\mathcal{H}_k^\epsilon$ in terms of these kernels and images allows us to identify Betti numbers at different grouping scales ϵ.

The quantum phase algorithm[38–40] allows one to decompose states in terms of the eigenvectors of an Hermitian matrix and to find the associated eigenvalues. Once the k-simplex states $|\psi\rangle_k^\epsilon$ have been constructed, the quantum phase algorithm allows one to decompose those states in terms of eigenvectors and eigenvalues of the boundary map. The boundary map is not Hermitian. We embed the boundary map $\tilde{\partial}_k$ into a Hermitian matrix B_k^ϵ defined by

$$B_k^\epsilon = \begin{pmatrix} 0 & \tilde{\partial}_k \\ \tilde{\partial}_k^\dagger & 0 \end{pmatrix}. \quad (3)$$

B_k^ϵ acts on the space $\mathcal{H}_{k-1}^\epsilon \oplus \mathcal{H}_k^\epsilon$. Note that B_k^ϵ is n-sparse: there are either k or $n-k$ entries per row. Similarly, define the full Hermitian boundary map to be

$$B^\epsilon = B_1^\epsilon \oplus B_2^\epsilon \oplus \ldots \oplus B_n^\epsilon. \quad (4)$$

B^ϵ is also n-sparse. Because $\tilde{\partial}_k \tilde{\partial}_{k+1} = 0$, we have $B^{\epsilon 2} = \Delta_0 \oplus \Delta_1 \oplus \ldots \oplus \Delta_n$, where $\Delta_k = \tilde{\partial}_k^\dagger \tilde{\partial}_k + \tilde{\partial}_{k+1} \tilde{\partial}_{k+1}^\dagger$ is the combinatorial Laplacian of the kth simplicial complex[22–24]. Because $(B^\epsilon)^2$ is the sum of the combinatorial Laplacians, B^ϵ is sometimes called the 'Dirac operator', since the original Dirac operator was the square root of the Laplacian. Explicit matrix forms of the Dirac operator and the combinatorial Laplacian are given in the Methods section. Hodge theory[9–12,14–24] implies that the kth homology group satisfies $\mathbf{H}_k = \mathrm{Ker}\,\tilde{\partial}_k / \mathrm{Image}_{k+1}\tilde{\partial}_{k+1} \cong \mathrm{Ker}\,\Delta_k$. The dimension of this kernel is the kth Betti number.

To find the dimension of the kernel, apply the quantum phase algorithm[38–40] to B^ϵ starting from the uniform mixture of simplices ρ^ϵ. The quantum phase algorithm decomposes this state into the eigenvectors of the combinatorial Laplacian, and identifies the corresponding eigenvalues. The probability of

yielding a particular eigenvalue is proportional to the dimension of the corresponding eigenspace. As above, classical algorithms for finding the eigenvalues and eigenvectors of the combinatorial Laplacians Δ_k, and calculating the dimension of the eigenspaces takes $O\left(\binom{n}{k}^2\right) \sim O(2^{2n})$ computational steps using sparse matrix diagonalization via Gaussian elimination or the Lanczos algorithm. On a quantum computer, however, the quantum phase algorithm[38–40] can project the simplex states $|\psi\rangle_k^\epsilon$ onto the eigenspaces of the Dirac operator B^ϵ and find corresponding eigenvalues to accuracy δ in time $O(n^5\delta^{-1}\zeta^{-1/2})$, where as above ζ is the accuracy to which we choose to construct the simplex state. The factor of n^5 arises because the quantum phase algorithm applied to an n-sparse matrix requires time n^3/δ^{-1}: the extra factor of n^2 arises because it takes time $O(k^2)$ to evaluate the projector P_k^ϵ onto the subspace of k simplices.

The algorithm also identifies the dimension of the eigenspaces of the Dirac operator and combinatorial Laplacian in time $O(n^5\delta^{-1}\zeta^{-1/2}\eta_\ell^{-1/2})$, where η_ℓ is equal to the dimension d_ℓ of the ℓth eigenspace divided by $|S|_k$, the dimension of the k-simplex space. The kth Betti number β_k is equal to the dimension of the kernel of Δ_k. The algorithm allows us to construct the full decomposition of the simplicial complex in terms of eigenvectors and eigenvalues of the combinatorial Laplacian, yielding useful geometric information such as harmonic forms. Monitoring how the eigenvalues and eigenspaces of the combinatorial Laplacian change as ϵ changes provides geometric information about how various topological features such as connected components, holes and voids come into existence and disappear as the grouping scale changes[16,17,44].

Discussion

This paper extended methods of quantum machine learning to topological data analysis. Homology is a powerful topological tool. The representatives of the homology classes for different k define the connected components of the simplicial complex, holes, voids and so on. The Betti numbers count the number of connected components, holes, voids and so on. Varying the simplicial scale ϵ and tracking how Betti numbers change as function of ϵ reveals how topological features come into existence and go away as the data is analysed at different length scales. Our algorithm also reveals how the structure of the eigenspaces and eigenvalues of the combinatorial Laplacian changes as a function of ϵ. This 'persistent geometry' reveals features of the data such as rate of change of harmonic forms over different simplicial scales.

The underlying methods of our quantum algorithms are similar to those in other big quantum data algorithms[19–21]. The primary difference between the topological and geometrical algorithms presented here, and algorithms for, for example, constructing clusters[19], principal components[20], and support vector machines[21], is that our topological algorithms require only a small qRAM of size $O(n^2)$. Consequently, even when the full qRAM resources are included in the accounting of the computational complexity of the algorithms, the topological algorithms require only an amount of computational resources polynomical in the number of data points, while the best existing classical algorithms for answering the same questions require exponential resources.

To recapitulate the steps of the algorithm: First, the quantum data is processed using standard techniques of quantum computation: distances between points are evaluated, simplices of neighbouring points are identified, and a simplicial complex is constructed. The simplicial complex depends on the grouping scale ϵ. We construct a quantum state that represents the filtration of the complex—the set of simplicial complexes, related by inclusion, for different ϵ. This quantum state contains exponentially fewer qubits than the number of bits required to describe the classical filtration of the complex. Second, we use the quantum phase algorithm[38–40] to calculate the eigenvalues and to construct the eigenspaces of the combinatorial Laplacian at each scale ϵ. The dimension of the kernel of the combinatorial Laplacian for k simplices is the kth Betti number. In addition, this construction gives us geometric information about the data set.

Classical algorithms for performing the full persistent homology over a space with n points over all scales k take time $O(2^{2n})$: there are 2^n possible simplices, and evaluating kernels and images of the boundary map via Gaussian elimination for sparse matrices takes time that goes as the square of the dimension of the space of simplices. By contrast, the quantum algorithm for constructing the Betti numbers and for decomposing the simplicial complex in terms of eigenvalues and eigenvectors of the combinatorial Laplacian takes time $O(n^5)$, compared with $O(2^{2n})$ for classical algorithms. The eigenvectors of the kernels of the combinatorial Laplacian are related to the representatives of the kth homology class via a boundary term. How to extend the quantum algorithms given here to construct the full barcode of persistent homology and to construct the representatives of the homology class directly is an open question. It would also be interesting to extend the quantum algorithmic methods developed here to further algebraic and combinatorial problems, for example, Morse theory.

Methods

Overview. In this section we provide further details of distance evaluation, simplex state construction, and the form of the Dirac operator and the combinatorial Laplacian.

State preparation and distance evaluation. Topological analysis of the data requires distances between data points. Assume that the data set contains n points together with the $n(n-1)/2$ distances between them. The data is stored in qRAM or qRAM[35–37], so that the algorithm can access the data in quantum parallel. The essential feature of a qRAM is that it preserves quantum coherence: the qRAM maps a quantum superposition of inputs $\sum_j \alpha_j|j\rangle|0\rangle$ to a quantum superposition of outputs $\sum_j \alpha_j|j\rangle|v_j\rangle$. Note that a quantum RAM is potentially significantly easier to construct than a full-blown quantum computer. The storage medium of a quantum RAM can be essentially classical: indeed, a single photon reflected off a compact disk encodes in its quantum state all the bits of information stored in the mirrors on the disk. In addition to a classical storage medium such as a CD, a qRAM contains quantum switches that can be opened in quantum superposition to access that information in quantum parallel. Each call to an N-bit qRAM requires $\log_2 N$ quantum operations. Quantum RAMS have been designed, and prototypes have been constructed[35–37]. In contrast to other big quantum data algorithms[31–33], the size of the qRAM required to perform topological and geometric analysis is relatively small: because the computational complexity of classical algorithms for persistent homology scales as $O(2^{2n})$, while the quantum algorithms require only $O(n^2)$ bits worth of qRAM, a significant quantum advantage could be obtained by a qRAM with hundreds to thousands of bits.

As an alternative to being presented with the pre-calculated distances, the data set could consist of n d-dimensional vectors $\{\vec{v}_j\}$ over the complex numbers, and we can use the qRAM to construct the distances $|\vec{v}_i - \vec{v}_j|$ between the ith and jth vectors[31]. Finally, the distances can be presented as the output of a quantum computation. In all cases, our quantum algorithms for topological and geometric analysis operate by accessing the distances in quantum parallel. Big quantum data analysis works by mapping each vector $\vec{v}_j$ to a quantum state $|v_j\rangle \in C^d$, and the entire database to a quantum state $(1/\sqrt{n})\sum_j |j\rangle|v_j\rangle \in C^n \otimes C^d$. A quantum RAM can be queried in quantum parallel: given an input state $|j\rangle|0\rangle$, it produces the output state $|j\rangle|v_j\rangle$, where $|v_j\rangle$ is normalized quantum state proportional to the vector $\vec{v}_j$. Such a quantum state can be encoded using $O(\log_2(nd))$ quantum bits, and $|\vec{v}_j|$ is the norm of the vector.

If we have not been given the $n(n-1)/2$ distances directly in qRAM, the next ingredient of the quantum algorithm is the ability to evaluate inner products and distances between vectors. In refs 20,31–33 it is shown how the access to vectors in quantum superposition: the ability to create the quantum states corresponding to the vectors translates into the ability to estimate

$\left|\vec{v}_i-\vec{v}_j\right|^2=2-\vec{v}_i^{\dagger}\vec{v}_j-\vec{v}_j^{\dagger}\vec{v}_i$. That is, we can construct a quantum circuit that takes as input the state $|i\rangle|j\rangle|0\rangle$ and produces as output the state $|i\rangle|j\rangle\left|\left|\vec{v}_i-\vec{v}_j\right|^2\right\rangle$, where the third register contains an estimate of the distance between $\vec{v}_i$ and $\vec{v}_j$. To estimate the distance to accuracy δ takes $O(\delta^{-1})$ quantum memory calls and $O(\delta^{-1}\ (\log_2(nd))^2)$ quantum operations. As with the qRAM, the circuit to evaluate distances operates in quantum parallel.

Simplex state construction. To elucidate the construction of the k-simplex states (1), we look more closely into the implementation of Grover's algorithm to understand when it succeeds in constructing the k-simplex state, and how it fails. Start from a superposition $n^{-1/2}\sum_k|k\rangle$ over all values of k. Performing simplex construction in parallel via Grover's algorithm with the membership function f_k^{ϵ} yields the full simplex state at scale ϵ:

$$|\Psi\rangle^{\epsilon}=\frac{1}{\sqrt{n}}\sum_k|k\rangle|\psi\rangle_k^{\epsilon}. \tag{5}$$

By adding ancillae as above, we can also construct the uniform mixture over all values of k and all k simplices: $\rho^{\epsilon}=(1/n)\sum_k|k\rangle\langle k|\otimes\rho_k^{\epsilon}$. More precisely, if we run the quantum search procedure for a time $\zeta^{-1/2}$, we will obtain the state

$$|\Psi\rangle_{\zeta}^{\epsilon}=\frac{1}{\sqrt{n}}\left(\sum_{k:\zeta_k^{\epsilon}\geq\zeta}|k\rangle|\psi\rangle_k^{\epsilon}+\sum_{k:\zeta_k^{\epsilon}<\zeta}|k\rangle|0\rangle\right) \tag{6}$$

that contains the simplex states $|\psi\rangle_k^{\epsilon}$ for which $\zeta_k^{\epsilon}\geq\zeta$ and which returns a null result $|0\rangle$ for the simplex states for which $\zeta_k^{\epsilon}<\zeta$. For small ϵ—where only a small fraction of all possible simplices lie within the complex—and fixed ζ, the simplex state $|\Psi\rangle_{\zeta}^{\epsilon}$ will contain the actual simplex states $|\psi\rangle_k^{\epsilon}$ only for small k. As ϵ becomes larger and larger, higher and higher k-simplex states enter the filtration and $|\Psi\rangle_{\zeta}^{\epsilon}$ will contain more and more of the k-simplex states.

Constructing the simplex state in quantum parallel at m different grouping scales ϵ_i yields the filtration state

$$|\Phi\rangle_{\zeta}=\frac{1}{\sqrt{mn}}\sum_i|\epsilon_i\rangle|\Psi\rangle_{\zeta}^{\epsilon}. \tag{7}$$

The filtration state $|\Phi\rangle_{\zeta}$ contains the entire filtration of the simplicial complex in quantum superposition. The quantum filtration state contains exponentially fewer quantum bits than the number of classical bits required to describe the classical filtration of the complex: $\log m$ qubits are required to register the grouping scale ϵ, and n qubits are required to label the simplices. $|\Phi\rangle_{\zeta}$ takes time $O(\zeta^{-1/2}n^2\log(m))$ to construct. By contrast, a classical description of the filtration of the simplicial complex requires $O(2^n)$ bits.

Explicit form of the Dirac operator and simplicial Laplacian. Here we present the full matrix form of the Dirac operator B^{ϵ} and the combinatorial Laplacian $(B^{\epsilon})^2$. The Dirac operator is

$$B^{\epsilon}=\begin{pmatrix}0&\tilde{\partial}_1&0&&&\\ \tilde{\partial}_1^{\dagger}&0&\tilde{\partial}_2&&\cdots&\\ 0&\tilde{\partial}_2^{\dagger}&0&&&\\ &&&\cdots&&\\ &&&0&\tilde{\partial}_{n-1}&0\\ &\cdots&&\tilde{\partial}_{n-1}^{\dagger}&0&\tilde{\partial}_n\\ &&&0&\tilde{\partial}_n^{\dagger}&0\end{pmatrix}, \tag{8}$$

where as above $\tilde{\partial}_k=P_{k-1}^{\epsilon}\partial_k P_k^{\epsilon}$ is the boundary map confined to the simplicial subspace $\mathcal{H}^{\epsilon}$. It is straightforward to verify that the Dirac operator is n-sparse.

The combinatorial Laplacian is obtained by squaring the Dirac operator:

$$(B^{\epsilon})^2=\begin{pmatrix}\tilde{\partial}_1\tilde{\partial}_1^{\dagger}&0&0&&&\\ 0&\tilde{\partial}_1^{\dagger}\tilde{\partial}_1+\tilde{\partial}_2\tilde{\partial}_2^{\dagger}&0&&\cdots&\\ 0&0&\tilde{\partial}_2^{\dagger}\tilde{\partial}_2+\tilde{\partial}_3\tilde{\partial}_3^{\dagger}&&&\\ &&\cdots&&&\\ &\cdots&&\tilde{\partial}_{n-1}^{\dagger}\tilde{\partial}_{n-1}+\tilde{\partial}_n\tilde{\partial}^{\dagger}&0&\\ &&&0&\tilde{\partial}_n^{\dagger}\tilde{\partial}_n&\end{pmatrix}. \tag{9}$$

The quantum algorithm operates by diagonalizing the Dirac operator.

References

1. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discret. Comput. Geom.* **33,** 249–274 (2005).
2. Robins, V. Towards computing homology from finite approximations. *Topol. Proc.* **24,** 503–532 (1999).
3. Frosini, P. & Landi, C. Size theory as a topological tool for computer vision. *Pattern Recognit. Image Anal.* **9,** 596–603 (1999).
4. Carlsson, G., Zomorodian, A., Collins, A. & Guibas, L. Persistence barcodes for shapes. *Int. J. Shape Model.* **11,** 149–188 (2005).
5. Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological persistence and simplification. *Discret. Comput. Geom.* **28,** 511–533 (2002).
6. Zomorodian, A. in *Algorithms and Theory of Computation Handbook* 2nd edn Ch. 3, section 2 (Chapman and Hall/CRC, 2009).
7. Chazal, F. & Lieutier, A. Stability and computation of topological invariants of solids in R^n. *Discret. Comput. Geom.* **37,** 601–617 (2007).
8. Cohen-Steiner, D., Edelsbrunner, H. & Harer, J. Stability of persistence diagrams. *Discret. Comput. Geom.* **37,** 103–120 (2007).
9. Basu, S. On bounding the Betti numbers and computing the euler characteristic of semi-algebraic sets. *Discret. Comput. Geom.* **22,** 1–18 (1999).
10. Basu, S. Different bounds on the different Betti numbers of semi-algebraic sets. *Discret. Comput. Geom.* **30,** 65–85 (2003).
11. Basu, S. Computing the top Betti numbers of semi-algebraic sets defined by quadratic inequalities in polynomial time. *Found. Comput. Math.* **8,** 45–80 (2008).
12. Basu, S. Algorithms in real algebraic geometry: a survey. Preprint at http://arxiv.org/abs/1409.1534 (2014).
13. Friedman, J. Computing Betti numbers via combinatorial Laplacians. in *Proceedings of the 28th Annual ACM Symposium on Theory of Computing*, 386–391 (Atlanta, Georgia, 1996).
14. Hodge, W. V. D. *The Theory and Applications of Harmonic Integrals* (Cambridge University Press, 1941).
15. Munkrees, J. R. *Elements of Algebraic Topology* (Benjamin/Cummings, 1984).
16. Butler, S. & Chung, F. Small spectral gap in the combinatorial Laplacian implies Hamiltonian. *Ann. Comb.* **13,** 403–412 (2010).
17. Maletić, S. & Rjković, M. Combinatorial Laplacian and entropy of simplicial complexes associated with complex networks. *Eur. Phys. J. Spec. Top.* **212,** 77–97 (2012).
18. Niyogi, P., Smale, S. & Weinberger, S. A topological view of unsupervised learning from noisy data. *SIAM J. Comput.* **40,** 646–663 (2011).
19. Kozlov, D. *Algorithms and Computation in Mathematics* Vol. 21 (Springer, 2008).
20. Ghrist, R. Barcodes: the persistent topology of data. *Bull. Am. Math. Soc.* **45,** 61–75 (2008).
21. Harker, S., Mischaikow, K., Mrozek, M. & Nanda, V. Discrete Morse theoretic algorithms for computing homology of complexes and maps. *Found. Comput. Math.* **14,** 151–184 (2014).
22. Mischaikow, K. & Nanda, V. Morse theory for filtrations and efficient computation of persistent homology. *Discret. Comput. Geom.* **50,** 330–353 (2013).
23. CHOMP. Computational homology project. http://chomp.rutgers.edu.
24. CAPD::RedHom: Reduction homology algorithms. http://redhom.ii.uj.edu.pl/.
25. Servedio, R. A. & Gortler, S. J. Equivalences and separations between quantum and classical learnability. *SIAM J. Comput.* **33,** 1067 (2004).
26. Hentschel, A. & Sanders, B. C. Machine learning for precise quantum measurement. *Phys. Rev. Lett.* **104,** 063603 (2010).
27. Neven, H., Denchev, V. S., Rose, G. & Macready, W. G. Training a large scale classifier with the quantum adiabatic algorithm. Preprint at http://arxiv.org/abs/0912.0779 (2009).
28. Pudenz, K. L. & Lidar, D. A. Quantum adiabatic machine learning. *Quantum Inf. Process* **12,** 2027 (2013).
29. Anguita, D., Ridella, S., Rivieccion, F. & Zunino, R. Quantum optimization for training support vector machines. *Neural Netw.* **16,** 763–770 (2003).
30. Aïmeur, E., Brassard, G. & Gambs, S. Quantum speed-up for unsupervised learning. *Mach. Lear.* **90,** 261–287 (2013).
31. Lloyd, S., Mohseni, M. & Rebentrost, P. Quantum algorithms for supervised and unsupervised machine learning. Preprint at http://arxiv.org/abs/1307.0411 (2013).
32. Rebentrost, P., Mohseni, M. & Lloyd, S. Quantum support vector machine for big feature and big data classification. *Phys. Rev. Lett.* **113,** 130503 (2014).
33. Lloyd, S., Mohseni, M. & Rebentrost, P. Quantum principal component analysis. *Nat. Phys.* **10,** 631–633 (2014).
34. Aaronson, S. Read the fine print. *Nat. Phys.* **11,** 291–293 (2015).
35. Giovannetti, V., Lloyd, S. & Maccone, L. Quantum random access memory. *Phys. Rev. Lett.* **100,** 160501 (2008).
36. Giovannetti, V., Lloyd, S. & Maccone, L. Architectures for a quantum random access memory. *Phys. Rev. A* **78,** 052310 (2008).
37. De Martini, F. *et al.* Experimental quantum private queries with linear optics. *Phys. Rev. A* **80,** 010302 (2009).
38. Yu. Kitaev, A., Shen, A. H. & Vyalyi, M. N. *Classical and Quantum Computation, Graduate Studies in Mathematics* Vol. 47 (publications of the American Mathematical Society, 2004).
39. Abrams, D. S. & Lloyd, S. A quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors. *Phys. Rev. Lett.* **83,** 5162–5165 (1999).

40. Nielsen, M. S. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge University Press, 2000).
41. Harrow, A. W., Hassidim, A. & Lloyd, S. Quantum algorithm for solving linear systems of equations. *Phys. Rev. Lett.* **15,** 150502 (2009).
42. Scheiblechner, P. On the complexity of deciding connectedness and computing Betti numbers of a complex algebraic variety. *J. Complex.* **23,** 359–379 (2007).
43. Kac, M. Can one hear the shape of a drum? *Am. Math. Mon.* **73,** 1–23 (1966).
44. Sadakane, K., Sugawara, N. & Tokuyama, T. Quantum computation in computational geometry. *Interdisc. Inf. Sci.* **8,** 129–136 (2002).

Acknowledgements

We thank Mario Rasetti for suggesting the topic of topological analysis of big data. We acknowledge helpful conversations with Patrick Rebentrost, Barbara Terhal and Francesco Vaccarino. S.L. was supported by ARO, AFOSR, DARPA and Jeffrey Epstein. P.Z. was supported by ARO MURI grant W911NF-11-1-0268 and by NSF grant PHY-969969.

Author contributions

All authors contributed to the problem formulation, quantum algorithm design and error analysis.

Additional information

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Lloyd, S. *et al.* Quantum algorithms for topological and geometric analysis of data. *Nat. Commun.* 7:10138 doi: 10.1038/ncomms10138 (2016).