# Estimating quantum speedups for lattice sieves

Martin R. Albrecht[1], Vlad Gheorghiu[2], Eamonn W. Postlethwaite[1], and John M. Schanck[2⋆]

[1] Information Security Group, Royal Holloway, University of London
[2] Institute for Quantum Computing, University of Waterloo, Canada

**Abstract.** Quantum variants of lattice sieve algorithms are routinely used to assess the security of lattice based cryptographic constructions. In this work we provide a heuristic, non-asymptotic, analysis of the cost of several algorithms for near neighbour search on high dimensional spheres. These algorithms are key components of lattice sieves. We design quantum circuits for near neighbour search algorithms and provide software that numerically optimises algorithm parameters according to various cost metrics. Using this software we estimate the cost of classical and quantum near neighbour search on spheres. For the most performant near neighbour search algorithm that we analyse we find a small quantum speedup in dimensions of cryptanalytic interest. Achieving this speedup requires several optimistic physical and algorithmic assumptions.

## 1 Introduction

Sieving algorithms for the shortest vector problem (SVP) in a lattice have received a great deal of attention recently [1, 2, 8, 17, 33, 40]. The attention mostly stems from lattice based cryptography, as many attacks on lattice based cryptographic constructions involve finding short lattice vectors [3, 36, 39].

Lattice based cryptography is thought to be secure against quantum adversaries. None of the known algorithms to solve SVP (to a small approximation factor) do so in subexponential time, but this is not to say that there is no gain to be had given a large quantum computer. Lattice sieve algorithms use near neighbour search (NNS) as a subroutine; near neighbour search algorithms use black box search as a subroutine; and Grover's quantum search algorithm [25] gives a square root improvement to the query complexity of black box search. A black box search that is expected to take $\Theta(N)$ queries on classical hardware will take $\Theta(\sqrt{N})$ queries on quantum hardware using Grover's algorithm.

⋆ The full version can be found at `https://eprint.iacr.org/2019/1161`. The research of MA was supported by EPSRC grants EP/S020330/1, EP/S02087X/1, by the European Union Horizon 2020 Research and Innovation Program Grant 780701 and Innovate UK grant AQuaSec; the research of EP was supported by the EPSRC and the UK government as part of the Centre for Doctoral Training in Cyber Security at Royal Holloway, University of London (EP/P009301/1). VG and JS were supported by NSERC and CIFAR. IQC is supported in part by the Government of Canada and the Province of Ontario.

Previous work has analysed the effect of quantum search on the query complexity of lattice sieves [34, 35]. Of course, one must implement the queries efficiently in order to realise the improvement in practice. Recent work has given concrete quantum resource estimates for the black box search problems involved in key recovery attacks on AES [23, 28] and preimage attacks on SHA-2 and SHA-3 [4]. In this work, we give explicit quantum circuits that implement the black box search subroutines of several quantum lattice sieves. Our quantum circuits are efficient enough to yield a cost improvement in dimensions of cryptanalytic interest. However, for the most performant sieve that we analyse the cost improvement is small and several barriers stand in the way of achieving it.

*Outline and Contributions.* We start with some preliminaries in Section 2. In particular, we discuss the "XOR and Population Count" operation (henceforth `popcount`), which is our primary optimisation target. The `popcount` operation is used to identify pairs of vectors that are likely to lie at a small angle to each other. It is typically less expensive than a full inner product computation.

In Section 3 we introduce and analyse a filtered quantum search procedure. We present our quantum circuit for `popcount` in Section 4. In Section 5 we provide a heuristic analysis of the probability that `popcount` successfully identifies pairs of vectors that are close to each other. This analysis may be of independent interest; previous work [2, 17] has relied largely on experimental data for choosing `popcount` parameters.

In Section 6, we rederive the overall cost of the NNS subroutines of three lattice sieves. Our cost analysis exposes the impact of the `popcount` parameters so that we can numerically optimise these in parallel with the sieve parameters. We have chosen to profile the Nguyen–Vidick sieve [40], the `bgj1` specialisation [2] of the Becker–Gama–Joux sieve [9], and the Becker–Ducas–Gama–Laarhoven sieve [8]. We have chosen these three sieves as they are, respectively, the earliest and most conceptually simple, the most performant yet implemented, and the fastest known asymptotically.

Finally, we optimise the cost of classical and quantum search under various cost metrics to produce Figure 2 of Section 7. We conclude by discussing barriers to obtaining the reported quantum advantages in NNS, the relationship between SVP and NNS, and future work. Both the data produced, and the source code used to compute it, are available at `https://github.com/jschanck/eprint-2019-1161`. We consider our software a contribution in its own right; it is documented, easily extensible and allows for the inclusion of new nearest neighbour search strategies and cost models.

*Interpretation.* Quantum computation seems to be more difficult than classical computation. As such, there will likely be some minimal dimension, a crossover point, below which classical sieves outperform quantum ones. Our estimates give non-trivial crossover points for the sieves we consider. Yet, our results do not rule out the relevance of quantum sieves to lattice cryptanalysis. The crossover points that we estimate are well below the dimensions commonly thought to achieve 128 bits of security against quantum adversaries. However, our initial

logical circuit level analysis (Figure 2, q: depth-width) is optimistic. It ignores the costs of quantum random access memory and quantum error correction.

To illustrate the potential impact of error correction, we apply a cost model developed by Gidney and Ekerå to our quantum circuits. The Gidney–Ekerå model was developed as part of a recent analysis of Shor's algorithm [20]. In the Gidney–Ekerå model, the crossover point for the NNS algorithm underlying the Becker–Ducas–Gama–Laarhoven sieve [8] is dimension 312. In this dimension, the classical and quantum variants both perform $2^{119.0}$ operations and need at least $2^{78.3}$ bits of (quantum accessible) random access memory. A large cost improvement is obtained asymptotically, but for cryptanalytically relevant dimensions the improvement is tenuous. Between dimensions 352 and 824 our estimate for the quantum cost grows from appoximately $2^{128}$ to approximately $2^{256}$. In dimension 352 this is an improvement of a factor of $2^{1.8}$ over our estimate for the classical cost. In dimension 824 the improvement is by a factor of $2^{14.4}$.

We caution that a memory constraint would significantly reduce the range of cryptanalytically relevant dimensions. For instance, an adversary with no more than $2^{128}$ bits of quantum accessible classical memory is limited to dimension 544 and below. In these dimensions we estimate a cost improvement of no more than a factor of $2^{13.6}$ at the logical circuit level and no more than $2^{7.1}$ in the Gidney–Ekerå metric.

A depth constraint would also reduce the range of cryptanalytically relevant dimensions. The quantum algorithms that we consider would be more severely affected by a depth constraint than their classical counterparts, due to the poor parallisability of Grover's algorithm.

*Acknowledgements.* We thank Léo Ducas for helpful discussions regarding ListDecodingSearch.

## 2 Preliminaries

### 2.1 Models of computation

We describe quantum algorithms as circuits using the Clifford+T gate set, but we augment this gate set with a table lookup operation (qRAM). We describe classical algorithms as programs for RAM machines (random access memory machines).

*Clifford+T+qRAM quantum circuits.* Quantum circuits can be described at the *logical layer*, wherein an array of $n$ qubits encodes a unit vector in $(\mathbb{C}^2)^{\otimes n}$, or at the *physical layer*, wherein the state space may be much larger. Ignoring qubit initialisation and measurement, a circuit is a sequence of unitary operations, one per unit time. Each unitary in the sequence is constructed by parallel composition of gates. At most one gate can be applied to each qubit per time

step. The Clifford+T gate set

$$\mathbf{H} = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix}1 & 0\\ 0 & i\end{pmatrix}, \quad \mathbf{CNOT} = \begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix}1 & 0\\ 0 & e^{i\pi/4}\end{pmatrix},$$

is commonly used to describe circuits at the logical layer due to its relationship with some quantum error correcting codes. This gate set is universal for quantum computation when combined with qubit initialisation (of $|0\rangle$ and $|1\rangle$ states) and measurement in the computational basis.

In addition to Clifford+T gates, we allow unit cost table lookups in the form of qRAM (quantum access to classical RAM). The difference between RAM and qRAM is that qRAM can construct arbitrary superpositions of table entries. Suppose that $(R_0, \ldots, R_{2^n-1})$ are registers of a classical RAM and that each register encodes an $\ell$ bit binary string. We allow our Clifford+T circuits access to these registers in the form of an $(n+\ell)$ qubit qRAM gate that enacts

$$\sum_{j=0}^{2^n-1} \alpha_j |j\rangle |x\rangle \stackrel{qRAM}{\longrightarrow} \sum_{j=0}^{2^n-1} \alpha_j |j\rangle |x \oplus R_j\rangle . \tag{1}$$

Here $\sum_j \alpha_j |j\rangle$ is a superposition of addresses and $x$ is an arbitrary $\ell$ bit string.

Quantum access to classical RAM is a powerful resource, and the algorithms we describe below fail to achieve an advantage over their classical counterparts when qRAM is not available. We discuss qRAM at greater length in Section 7.

*RAM machines.* We describe classical algorithms in terms of random access memory machines. For comparability with the Clifford+T gate set, we will work with a limited instruction set, e.g. {NOT, AND, OR, XOR, LOAD, STORE}. For comparability with qRAM, LOAD and STORE act on $\ell$ bit registers.

*Cost.* The cost of a RAM program is the number of instructions that it performs. One can similarly define the *gate cost* of a quantum circuit to be the number of gates that it performs. Both metrics are reasonable in isolation, but it is not clear how one should compare the two. Jaques and Schanck recommend that quantum circuits be assigned a cost in the unit of RAM instructions to account for the role that classical computers play in dispatching gates to quantum memories [29]. They also recommend that the identity gate be assigned unit cost to account for error correction. The *depth-width cost* of a quantum circuit is the total number of gate operations that it performs when one includes identity gates in the count.

### 2.2 Black box search

A predicate on $\{0, 1, \ldots, N-1\}$ is a function $f : \{0, 1, \ldots, N-1\} \to \{0, 1\}$. The kernel, or set of roots, of $f$ is $\mathrm{Ker}(f) = \{x : f(x) = 0\}$. We write $|f|$ for $|\mathrm{Ker}(f)|$. A black box search algorithm finds a root of a predicate without exploiting any

structure present in the description of the predicate itself. Of course, black box search algorithms can be applied when structure is known, and we will often use structure such as "$f$ has $M$ roots" or "$f$ is expected to have no more than $M$ roots" in our analyses. We will also use the fact that the set of predicates on any given finite set can be viewed as a Boolean algebra. We write $f \cup g$ for the predicate with kernel $\mathrm{Ker}(f) \cup \mathrm{Ker}(g)$ and $f \cap g$ for the predicate with kernel $\mathrm{Ker}(f) \cap \mathrm{Ker}(g)$.

*Exhaustive search.* An exhaustive search evaluates $f(0)$, $f(1)$, $f(2)$, and so on until a root of $f$ is found. The order does not matter so long as each element of the search space is queried at most once. If $f$ is a uniformly random predicate with $M$ roots, then this process has probability $1 - \binom{N-M}{j}/\binom{N}{j} \geq 1 - (1 - M/N)^j$ of finding a root during $j$ evaluations of $f$. This is true even if $M$ is not known.

*Filtered search.* If $f$ is expensive to evaluate, we may try to decrease the cost of exhaustive search by applying a search filter. We say that a predicate $g$ is a filter for $f$ if $f \neq g$ and $|f \cap g| \geq 1$. We say that $g$ recognises $f$ with a false positive rate of

$$\rho_f(g) = 1 - \frac{|f \cap g|}{|g|},$$

and a false negative rate of

$$\eta_f(g) = 1 - \frac{|f \cap g|}{|f|}.$$

A filtered search evaluates $g(0), f(0), g(1), f(1), g(2), f(2)$, and so on until a root of $f \cap g$ is found. The evaluation of $f(i)$ can be skipped when $i$ is not a root of $g$, which may reduce the cost of filtered search below that of exhaustive search.

*Quantum search.* Grover's quantum search algorithm is a black box search algorithm that provides a quadratic advantage over exhaustive search in terms of query complexity. Suppose that $f$ is a predicate with $M$ roots. Let $\mathbf{D}$ be any unitary transformation that maps $|0\rangle$ to $\frac{1}{\sqrt{N}} \sum_i |i\rangle$, let $\mathbf{R}_0 = \mathbf{I}_N - 2|0\rangle\langle 0|$ and let $\mathbf{R}_f$ be the unitary $|x\rangle \mapsto (-1)^{f(x)}|x\rangle$. Measuring $\mathbf{D}|0\rangle$ yields a root of $f$ with probability $M/N$. Grover's quantum search algorithm amplifies this to probability $\approx 1$ by repeatedly applying the unitary $\mathbf{G}(f) = \mathbf{D}\mathbf{R}_0\mathbf{D}^{-1}\mathbf{R}_f$ [25]. Suppose that $j$ repetitions are applied. The analysis in [25] shows that measuring the state $\mathbf{G}(f)^j\mathbf{D}|0\rangle$ yields a root of $f$ with probability $\sin^2((2j+1)\cdot\theta)$ where $\sin^2(\theta) = M/N$. Assuming $M \ll N$, the probability of success is maximised at $j \approx \frac{\pi}{4}\sqrt{N/M}$ iterations. Boyer, Brassard, Høyer, and Tapp (BBHT) show that a constant success probability can be obtained after $O(\sqrt{N/M})$ iterations.

The same complexity can be obtained when $M$ is not known. One simply runs the algorithm repeatedly with $j$ chosen uniformly from successively larger intervals. The following lemma contains the core observation.

**Lemma 1 (Lemma 2 of [12]).** *Suppose that measuring* $\mathbf{D}|0\rangle$ *would yield a root of* $f$ *with probability* $\sin^2(\theta)$*. Fix a positive integer* $m$*. Let* $j$ *be chosen uniformly from* $\{0, \ldots, m-1\}$*. The expected probability that measuring* $\mathbf{G}(f)^j\mathbf{D}|0\rangle$ *yields a root of* $f$ *is* $\frac{1}{m}\sum_{j=0}^{m-1}\sin^2((2j+1)\cdot\theta) = \frac{1}{2} - \frac{\sin(4m\theta)}{4m\,\sin(2\theta)}$*. If* $m > 1/\sin(2\theta)$ *then this quantity is at least* $1/4$*.*

The complete strategy is made precise by [12, Theorem 3].

*Amplitude amplification.* Brassard, Høyer, Mosca, and Tapp observed that the $\mathbf{D}$ subroutine of Grover's algorithm can be replaced with any algorithm that finds a root of $f$ with positive probability [13]. This generalisation of Grover's algorithm is called amplitude amplification. Let $\mathbf{A}$ be a quantum algorithm that makes no measurements and let $p$ be the probability that measuring $\mathbf{A}|0\rangle$ yields a root of $f$. Let $\mathbf{G}(\mathbf{A}, f) = \mathbf{A}\mathbf{R}_0\mathbf{A}^{-1}\mathbf{R}_f$, where $\mathbf{R}_0$ and $\mathbf{R}_f$ are as in Grover's algorithm. Let $\theta$ be such that $\sin^2(\theta) = p$. Suppose that $j$ iterations of $\mathbf{G}(\mathbf{A}, f)$ are applied to $\mathbf{A}|0\rangle$. The analysis in [13] shows that measuring the state $\mathbf{G}(\mathbf{A}, f)^j\mathbf{A}|0\rangle$ yields a root of $f$ with probability $\sin^2((2j+1)\cdot\theta)$. The BBHT strategy for handling an unknown number of roots generalises to an unknown $p$.

### 2.3 Lattice sieving and near neighbour search on the sphere

A Euclidean lattice of rank $m$ and dimension $d$ is an abelian group generated by integer sums of $m \leq d$ linearly independent vectors in $\mathbb{R}^d$. In this paper we only consider full rank lattices, i.e. $m = d$. The shortest vector problem in a lattice $\Lambda$ is the problem of finding a non-zero $v \in \Lambda$ of minimal Euclidean norm. Norms in this work are Euclidean and denoted $\|\cdot\|$. The angular distance of $u, v \in \mathbb{R}^d$ is denoted $\theta(u, v) = \arccos\left(\langle u, v\rangle / \|u\|\|v\|\right)$, $\arccos(x) \in [0, \pi]$.

A lattice sieve takes as input a list of lattice points, $L \subset \Lambda$, and searches for integer combinations of these points that are short. If the initial list is sufficiently large, SVP can be solved by performing this process recursively. Each point in the initial list can be sampled at a cost polynomial in $d$ [31]. Hence the initial list can be sampled at a cost of $|L|^{1+o(1)}$.

Sieves that combine $k$ points at a time are called $k$-sieves. The sieves that we consider in this paper are 2-sieves. They take integer combinations of the form $u \pm v$ with $u, v \in L$ and $u \neq \pm v$. If $\|u \pm v\| \geq \max\{\|u\|, \|v\|\}$ then we say that $(u, v)$ is a reduced pair, else it is a reducible pair.

We analyse 2-sieves under the heuristic that the points in $L$ are independent and identically distributed (i.i.d.) uniformly in a thin spherical shell. This heuristic was introduced by Nguyen and Vidick in [40]. As a further simplification, we assume that the shell is very thin and normalise such that $L \subset \mathcal{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$. As such, $(u, v)$ are reducible if and only if $\theta(u, v) < \pi/3$. The `popcount` filter, introduced in Section 2.4, acts as a first approximation to $\theta(\cdot, \cdot)$.

When we model $L$ as a subset of $\mathcal{S}^{d-1}$, we can translate some lattice sieves into the language of (angular) near neighbour search on the sphere. For example,

the Nguyen–Vidick sieve [40], which checks all pairs in $L$ for reducibility, becomes[1] Algorithm 1 with $\theta = \pi/3$.

**Algorithm 1** AllPairSearch

**Input:** A list $L = (v_1, v_2, \dots v_N) \subset \mathcal{S}^{d-1}$ of $N$ points. Parameter $\theta \in (0, \pi/2)$.
**Output:** A list of pairs $(u, v) \in L \times L$ with $\theta(u, v) \leq \theta$.

**function** AllPairSearch($L; \theta$)
  $L' \leftarrow \emptyset$
  **for** $1 \leq i < N$ **do**
    $L_i \leftarrow (v_{i+1}, \dots, v_N)$
    Search $L_i$ for any number of $u$ that satisfy $\theta(u, v_i) \leq \theta$.
    For each such $u$ found, add $(u, v_i)$ to $L'$.
    If $|L'| \geq N$, **return** $L'$.
  **return** $L'$

### 2.4 The popcount filter

Charikar's locality sensitive hashing (LSH) scheme [15] is a family of hash functions $\mathcal{H}$, defined on $\mathcal{S}^{d-1}$, for which

$$\Pr_{h \leftarrow \mathcal{H}}[h(u) = h(v)] = 1 - \frac{\theta(u, v)}{\pi}. \tag{2}$$

The hash function family is defined by

$$\mathcal{H} = \left\{ u \mapsto \operatorname{sgn}(\langle r, u \rangle) : r \in \mathcal{S}^{d-1} \right\},$$

where $\operatorname{sgn}(x) = 1$ if $x \geq 0$ and $\operatorname{sgn}(x) = 0$ if $x < 0$. Equation 2 follows from the fact that $\theta(u, v)/\pi$ is the probability that uniformly random $u$ and $v$ lie in opposite hemispheres.

Charikar observed that one can estimate $\theta(u, v)/\pi$ by choosing a random hash function $h = (h_1, \dots, h_n) \in \mathcal{H}^n$ and measuring the Hamming distance between $h(u) = (h_1(u), \dots, h_n(u))$ and $h(v) = (h_1(v), \dots, h_n(v))$. Each bit $h_i(u) \oplus h_i(v)$ is Bernoulli distributed with parameter $p = \theta(u, v)/\pi$. In the limit of large $n$, the normalised Hamming weight $wt(h(u) \oplus h(v))/n$ converges to a normal distribution with mean $p$ and standard deviation $\sqrt{p(1-p)/n}$.

In the sieving literature, the process of filtering a $\theta(\cdot, \cdot)$ test using a threshold on the value of $wt(h(u) \oplus h(v))$ is known as the "XOR and population count

[1] This is slightly imprecise. The analogy with the Nguyen–Vidick sieve is completed only when Algorithm 1 is wrapped in a procedure that takes each $(u, v) \in L'$ and maps it to $(u \pm v)/\|u \pm v\|$, and then recurses.

trick" [2, 17, 18]. Functions in $\mathcal{H}^n$ are also used in Laarhoven's HashSieve [33]. We write $\texttt{popcount}_{k,n}(u, v; h)$ for a search filter of this type

$$\texttt{popcount}_{k,n}(u, v; h) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} h_i(u) \oplus h_i(v) \leq k, \\ 1 & \text{otherwise.} \end{cases}$$

When the $n$ hash functions are fixed we write $\texttt{popcount}_{k,n}(u, v)$. The threshold, $k$, is chosen based on the desired false positive and false negative rates. Heuristically, if one's goal is to detect points at angle at most $\theta$, one should take $k/n \approx \theta/\pi$. If $k/n \ll \theta/\pi$ then the false negative rate will be large, and many neighbouring pairs will be missed. An important consequence of missing potential reductions is that the $N$ required to iterate Algorithms 1, 3, 4 increases. In Section 6 this increase is captured in the quantity $\ell(k, n)$. If $k/n \gg \theta/\pi$ then the false positive rate will be large, and the full inner product test will be applied often. We calculate these false positive and negative rates in Section 5. These calculations and the fact that `popcount` is significantly cheaper than an inner product makes `popcount` a good candidate for use as a filter under the techniques of Section 2.2. Furthermore it is the filter used in the most performant sieves to date [2, 17].

### 2.5 Geometric figures on the sphere

Our analysis of the `popcount` filter requires some basic facts about the size of some geometric figures on the sphere. We measure the volume of subsets of $\mathcal{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$ using the $(d-1)$ dimensional spherical probability measure[2] $\mu^{d-1}$. The spherical cap of angle $\theta$ about $u \in \mathcal{S}^{d-1}$ is $\mathcal{C}^{d-1}(u, \theta) = \{v \in \mathcal{S}^{d-1} : \theta(u, v) \leq \theta\}$. The measure of a spherical cap is

$$C_d(u, \theta) := \mu^{d-1}(\mathcal{C}^{d-1}(u, \theta)) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})} \int_0^\theta \sin^{d-2}(t)\ \mathrm{d}t.$$

We will often interpret $C_d(u, \theta)$ as the probability that $v$ drawn uniformly from $\mathcal{S}^{d-1}$ satisfies $\theta(u, v) \leq \theta$. We denote the density of the event $\theta(u, v) = \theta$ by

$$A_d(u, \theta) := \frac{\partial}{\partial \theta} C_d(u, \theta) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})} \sin^{d-2}(\theta).$$

Note that $C_d(u, \theta)$ does not depend on $u$, so we may write $C_d(\theta)$ and $A_d(\theta)$ without ambiguity. The wedge formed by the intersection of two caps is $\mathcal{W}^{d-1}(u, \theta_u, v, \theta_v) = \mathcal{C}^{d-1}(u, \theta_u) \cap \mathcal{C}^{d-1}(v, \theta_v)$. The measure of a wedge only depends on $\theta = \theta(u, v)$, $\theta_u$, and $\theta_v$, so we denote it

$$W_d(\theta, \theta_u, \theta_v) = \mu^{d-1}(\mathcal{W}^{d-1}(u, \theta_u, v, \theta_v)).$$

We will often interpret $W_d(\theta, \theta_u, \theta_v)$ as the probability that $w$ drawn uniformly from $\mathcal{S}^{d-1}$ satisfies $\theta(u, w) \leq \theta_u$ and $\theta(v, w) \leq \theta_v$. Note that $\theta \geq \theta_u + \theta_v \Rightarrow W_d(\theta, \theta_u, \theta_v) = 0$. An integral representation of $W_d(\theta, \theta_u, \theta_v)$ is given in Appendix A of the full version.

[2] By "probability measure" we mean that $\mu^{d-1}(\mathcal{S}^{d-1}) = 1$.

## 3 Filtered quantum search

A filter can reduce the cost of a search because a classical computer can branch to avoid evaluating an expensive predicate. A quantum circuit cannot branch inside a Grover search in this way. Nevertheless, a filter can be used to reduce the cost of a quantum search.

The idea is to apply amplitude amplification to a Grover search. The inner Grover search prepares the uniform superposition over roots of the filter, $g$. The outer amplitude amplification searches for a root of $f$ among the roots of $g$. We present pseudocode for this strategy in Algorithm 2.

If $|g|$ and $|f \cap g|$ are known, then we can choose the number of iterations of the inner Grover search and the outer amplitude amplification optimally. When these quantities are not known, we can attempt to guess them as in the BBHT algorithm. In our applications, we have some information about $|g|$ and $|f \cap g|$, which we can use to fine-tune a BBHT-like strategy.

Proposition 1 gives the cost of Algorithm 2 when we know *1.* a lower bound, $Q$, on the size of $|f \cap g|$, and *2.* the value of $|g|$ up to relative error $\gamma$. In essence, when a filter with a low false positive rate is used to search a space with few true positives, Algorithm 2 can be tuned such that it finds a root of $f$ with probability at least $1/14$ and at a cost of roughly $\frac{\gamma}{2}\sqrt{N/Q}$ iterations of $\mathbf{G}(g)$.

**Algorithm 2** FilteredQuantumSearch

**Input:** A predicate $f$ and a filter $g$ defined on $\{0, \ldots, N-1\}$. Integer parameters $m_1$ and $m_2$.

**Output:** A root of $f$ or $\perp$.

**function** FilteredQuantumSearch($f, g; m_1, m_2$)
  Sample integers $j$ and $k$ with $0 \leq j < m_1$ and $0 \leq k < m_2$ uniformly at random.
  Let $\mathbf{A}_j = \mathbf{G}(g)^j\mathbf{D}$.
  Let $\mathbf{B}_k = \mathbf{G}(\mathbf{A}_j, f \cap g)^k$.
  Prepare the state $|\psi\rangle = \mathbf{B}_k\mathbf{A}_j|0\rangle$.
  Let $r$ be the result of measuring $|\psi\rangle$ in the computational basis.
  **if** $f(r) = 0$ **then**
    **return** $r$
  **return** $\perp$

If we know that the the inner Grover search succeeds with probability $x < 1$, we can compensate with a factor of $\sqrt{1/x}$ more iterations of the outer amplitude amplification. We do not know $x$. However, in our applications, we do know that the value of $\theta$ for which $\sin^2(\theta) = |g|/N$ will be fairly small, e.g. $\theta < 1/10$. The following technical lemma shows that, when $\theta$ is small, we may assume that $x = 1/5$ with little impact on the overall cost of the search.

Let $j$ and $\mathbf{A}_j$ be as in Algorithm 2. Let $p_\theta(j)$ be the probability that measuring $\mathbf{A}_j|0\rangle$ would yield a root of $g$. For any $x \in (0, 1)$, there is some probability $q_x(m_1)$ that the choice of $j$ is insufficient, i.e. that $p_\theta(j) < x$. We expect to repeat Algorithm 2 a total of $(1 - q_x(m_1))^{-1}$ times to avoid this type of failure.

**Lemma 2.** *Fix $\theta \in [0, \pi/2]$ and $x \in [0, 1)$. Let $p_\theta, q_x : \mathbb{R} \to \mathbb{R}$ be defined by $p_\theta(j) = \sin^2((2j+1) \cdot \theta)$ and $q_x(m) = \frac{1}{m} |\{j \in \mathbb{Z} : 0 \leq j < m, p_\theta(j) < x\}|$. If $m > \frac{\pi}{4\theta}$, then*

$$q_x(m) < \frac{3 \arcsin(\sqrt{x})}{\pi - \arcsin(\sqrt{x})} + \frac{6\theta}{\pi}.$$

*Proof.* Observe that $p_\theta(j) < x$ when $|(2j+1)\theta \bmod \pi| < \arcsin(\sqrt{x})$. Let $I_0$ be the interval $[0, \arcsin(\sqrt{x}))$. For integers $t \geq 1$ let $I_t = (t\pi - \arcsin(\sqrt{x}), t\pi + \arcsin(\sqrt{x}))$. Let $c = c(m)$ be the largest integer for which $[0, (2m-1) \cdot \theta)$ intersects $I_c$. The quantity $mq_x(m)$ counts the number of non-negative integers $i < m$ for which $(2i+1) \cdot \theta$ lies in $I_0 \cup I_1 \cup \cdots \cup I_c$. This is no more than $(c+1) + \lfloor (2c+1) \arcsin(\sqrt{x})/(2\theta) \rfloor$. It follows that $q_x(m) < (c+1)/m + (2c+1) \arcsin(\sqrt{x})/2m\theta$. Note that $2m\theta > (2m-1)\theta > c\pi - \arcsin(\sqrt{x})$ and $(c+1)/m < 2\theta/\pi + 1/m$. Hence $q_x(m) < (2c+1) \arcsin(\sqrt{x})/(c\pi - \arcsin(\sqrt{x})) + 2\theta/\pi + 1/m$. Moreover, $q_x(m) > q_x(m-1)$ when $(2m-1) \cdot \theta$ lies in $I_c$, and $q_x(m) < q_x(m-1)$ otherwise. The upper bound on $q_x(m)$ that we have derived is decreasing as a function of $c$. Hence the claim holds when $c \geq 1$. Finally, when $m = \frac{\pi}{4\theta}$ and $c = 0$ we have $q_x(m) < 2 \arcsin(\sqrt{x})/\pi + 4\theta/\pi$ and $q_x(m)$ is decreasing until $c = 1$. □

There are situations in which filtering is not effective, e.g. when the false positive rate of $g$ is very high, when evaluting $g$ is not much less expensive than evaluating $f$, or when $f$ has a very large number of roots. In these cases, other algorithms will outperform Algorithm 2. We remark on these below. Proposition 1 optimises the choice of $m_1$ and $m_2$ in Algorithm 2 for a large class of filters that are typical of our applications.

**Proposition 1.** *Suppose that $f$ and $g$ are predicates on a domain of size $N$ and that $g$ is a filter for $f$. Let $Q \in \mathbb{R}$ be such that $1 \leq Q \leq |f \cap g|$. Let $P$ and $\gamma$ be real numbers such that $P/\gamma \leq |g| \leq \gamma P$. If $\gamma P/N < 1/100$ and $\gamma Q/P < 1/4$, then there are parameters $m_1$ and $m_2$ for Algorithm 2 such that Algorithm 2 finds a root of $f$ with probability at least $1/14$ and has a cost that is dominated by $\approx \frac{\gamma}{2}\sqrt{N/Q}$ times the cost of $\mathbf{G}(g)$ or by $\approx \frac{2}{3}\sqrt{\gamma P/Q}$ times the cost of $\mathbf{R}_{f \cap g}$.*

*Proof.* Fix $x \in (0, 1)$. We will analyse Algorithm 2 with respect to the parameters $m_1 = \left\lceil \frac{\pi}{4}\sqrt{\gamma N/P} \right\rceil$ and $m_2 = \left\lceil \sqrt{\gamma P/3xQ} \right\rceil$. Let $\theta_g$ be such that $\sin^2(\theta_g) = |g|/N$. Let $j$ and $k$ be chosen as in Algorithm 2. Let $p = p_{\theta_g}(j)$ and $q = q_x(m_1)$ be defined as in Lemma 2. Note that since $|g|/N < \gamma P/N < 1/100$ we can use $6\theta_g/\pi < 1/5$ in applying Lemma 2. Let $\theta_h(j)$ be such that $\sin^2(\theta_h(j)) = p \cdot |f \cap g| / |g|$. With probability at least $1 - q$ we have $p \geq x$, which implies that $\sin(\theta_h(j)) \geq \sqrt{xQ/\gamma P}$. Since $\gamma Q/P < 1/4 \Rightarrow \sin^2(\theta_h(j)) < 1/4$, then $\cos(\theta_h(j)) > \sqrt{3/4}$. Thus $1/\sin(2\theta_h(j)) < \sqrt{\frac{\gamma P}{3xQ}} \leq m_2$. By Lemma 1 measuring $\mathbf{G}(\mathbf{A}_j, f \cap g)^k \mathbf{A}_j |0\rangle$ yields a root of $f \cap g$ with probability at least $1/4$. It follows that Algorithm 2 succeeds with probability at least $(1-q)/4$.

The algorithm evaluates $\mathbf{G}(g)$ exactly $k \cdot j + 1$ times and evaluates $\mathbf{G}(g)^{-1}$ exactly $k \cdot j$ times. The expected value of $2kj + 1$ is $c_1(x) \cdot \gamma \cdot \sqrt{N/Q}$ where $c_1(x) \approx (\pi/8)/\sqrt{3x}$. Likewise the algorithm evaluates $\mathbf{R}_{f \cap g}$ exactly $k$ times, which

is $c_2(x) \cdot \sqrt{\gamma P/Q}$ in expectation where $c_2(x) \approx (1/2)/\sqrt{3x}$. Taking $x = 1/5$, and applying the upper bound on $q_x(m_1)$ from Lemma 2, we have $(1 - q_x(m_1))/4 \geq 1/14$, $c_1(x) \approx 1/2$ and $c_2(x) \approx 2/3$. □

*Remark 1.* When $\gamma P/N \geq 1/100$ or $\gamma Q/P \geq 1/4$ there are better algorithms. If both inequalities hold then classical search finds a root of $f$ quickly. If $\gamma Q/P \geq 1/4$ then finding a root of $f$ is not much harder than finding a root of $g$, so one can search on $g$ directly. If $\gamma P/N \geq 1/100$ then the filter has little effect and one can search on $f$ directly.

*Remark 2.* It is helpful to understand when we can ignore the cost of $\mathbf{R}_{f \cap g}$ in Proposition 1. Roughly speaking, if evaluating $f$ is $c$ times more expensive than evaluating $g$, then the cost of calls to $\mathbf{G}(g)$ will dominate when $N > c^2 |g|$. In a classical filtered search the cost of evaluating $g$ dominates when $N > c |g|$.

## 4 Circuits for `popcount`

Consider a program for $\texttt{popcount}_{k,n}(u, v)$. This program loads $u$ and $v$ from specified memory addresses, computes $h(u)$ and $h(v)$, computes the Hamming weight of $h(u) \oplus h(v)$, and checks whether it is less than or equal to $k$. Recall $h(u)$ is defined by $n$ inner products. If the `popcount` procedure is executed many times for each $u$, then it may be reasonable to compute $h(u)$ once and store it in memory. Moreover, if $u$ is fixed for many sequential calls to the procedure, then it may be reasonable to cache $h(u)$ between calls. The algorithms that we consider in Section 6 use both of these optimisations.

In this section we describe RAM programs and quantum circuits that compute $\texttt{popcount}_{k,n}(u, \cdot)$ for a fixed $u$. These circuits have the value of $h(u)$ hard-coded. They load $h(v)$ from memory, compute the Hamming weight of $h(u) \oplus h(v)$, and check whether the Hamming weight is less than or equal to $k$. We ignore the initial, one time, cost of computing $h(u)$ and $h(v)$.

### 4.1 Quantum circuit for `popcount`

Loading $h(v)$ costs a single qRAM gate. Computing $h(u) \oplus h(v)$ can then be done in-place using a sequence of $\mathbf{X}$ gates that encode $h(u)$. The bulk of the effort is in computing the Hamming weight. For that we use a tree of in-place adders. The final comparison is also computed with an adder, although only one bit of the output is needed. See Figure 1 for a full description of the circuit.

We use the Cuccaro–Draper–Kutin–Petrie adder [16], with "incoming carry" inputs, to compute the Hamming weight. We argue in favour of this choice of adder in Appendix C of the full version. We use the Häner–Roetteler–Svore [26] carry bit circuit for implementing the comparison.

We will later use `popcount` within filtered quantum searches by defining predicates of the form $g(i) = \texttt{popcount}_{k,n}(u, v_i)$, $i \in \{1, \ldots, N\}$. To simplify that later discussion, we cost the entire Grover iteration $\mathbf{G}(g) = \mathbf{D}\mathbf{R}_0\mathbf{D}^{-1}\mathbf{R}_g$ here. In Appendix B of the full version we introduce the (possibly multiply controlled)

Toffoli gate and discuss the Toffoli count for $\mathbf{G}(g)$, which in turn gives the $\mathbf{T}$ count for $\mathbf{G}(g)$.

*The cost of* $\mathbf{R}_g$. The $\mathbf{R}_g$ subroutine is computed by running the `popcount` circuit in Figure 1 and then uncomputing the addition tree and $\mathbf{X}$ gates. The circuit uses in-place $i$ bit adders[3] for $i \in \{1, \ldots, \ell - 1\}$. The width of the circuit is given in Appendix B of the full version. The depth of the circuit is

$$\text{depth} = 2 + d(\text{CARRY}) + \sum_{i=1}^{\ell-1} 2 \cdot d(\text{ADD}_i), \tag{3}$$

where $d(\cdot)$ denotes the depth of its argument. The factor of 2 accounts for uncomputation of the $\text{ADD}_i$ circuits. The CARRY circuit is only cost once as the carry bit is computed directly into the $|-\rangle$ state during the CARRY circuit itself. The summand 2 accounts for the $\mathbf{X}$ gates used to compute, and later uncompute, $h(u) \oplus h(v)$.

*The cost of* $\mathbf{D}\mathbf{R}_0\mathbf{D}^{-1}$. Recall that $\mathbf{D}$ can be any circuit that maps $|0\rangle$ to the uniform distribution on the domain of the search predicate. While there is no serious difficulty in sampling from the uniform distribution on $\{0, \ldots, N-1\}$ for any integer $N$, when costing the circuit we assume that $N$ is a power of two. In this case $\mathbf{D}$ is simply $\log_2 N$ parallel $\mathbf{H}$ gates. The reflection $\mathbf{R}_0$ is implemented as a multiply controlled Toffoli gate that targets an ancilla initialised in the $|-\rangle$ state. We use Maslov's multiply controlled Toffoli from [37]. The depth and width of $\mathbf{D}\mathbf{R}_0\mathbf{D}^{-1}$ are both $O(\log N)$; our software calculates the exact value.

### 4.2 RAM program for `popcount`

Recall that we use a RAM instruction set that consists of simple bit operations and table lookups. A Boolean circuit for `popcount` is schematically similar to Figure 1. Let $\ell = \lceil \log_2 n \rceil$. Loading $h(v)$ has cost 1. Computing $h(v) \oplus h(w)$ takes $n$ XOR instructions and has depth 1. Following [41, Table. II], with $c_{FA} = 5$ the number of instructions in a full adder, $(n - \ell - 1)c_{FA} + \ell$ lower bounds the instruction cost of computing the Hamming weight and comparing it with a fixed $k$. This has depth $(\ell - 1)(\delta_{\text{sum}} + \delta_{\text{carry}}) + 1$. We assume $\delta_{\text{sum}} = \delta_{\text{carry}} = 1$. Thus, the overall instruction count is $6n - 4\ell - 5$ and the overall depth is $2\ell$.

### 4.3 Cost of inner products

The optimal `popcount` parameters will depend on the cost of a computing an inner product in dimension $d$. The cost of one inner product is amortised over many `popcounts`, and a small change in the `popcount` parameters will quickly

[3] An in-place $i$ bit quantum adder takes two $i$ bit inputs, initialises an ancilla qubit in the $|0\rangle$ state, and returns the addition result in an $i + 1$ bit register that includes the new ancilla and overlaps with $i$ bits of the input.

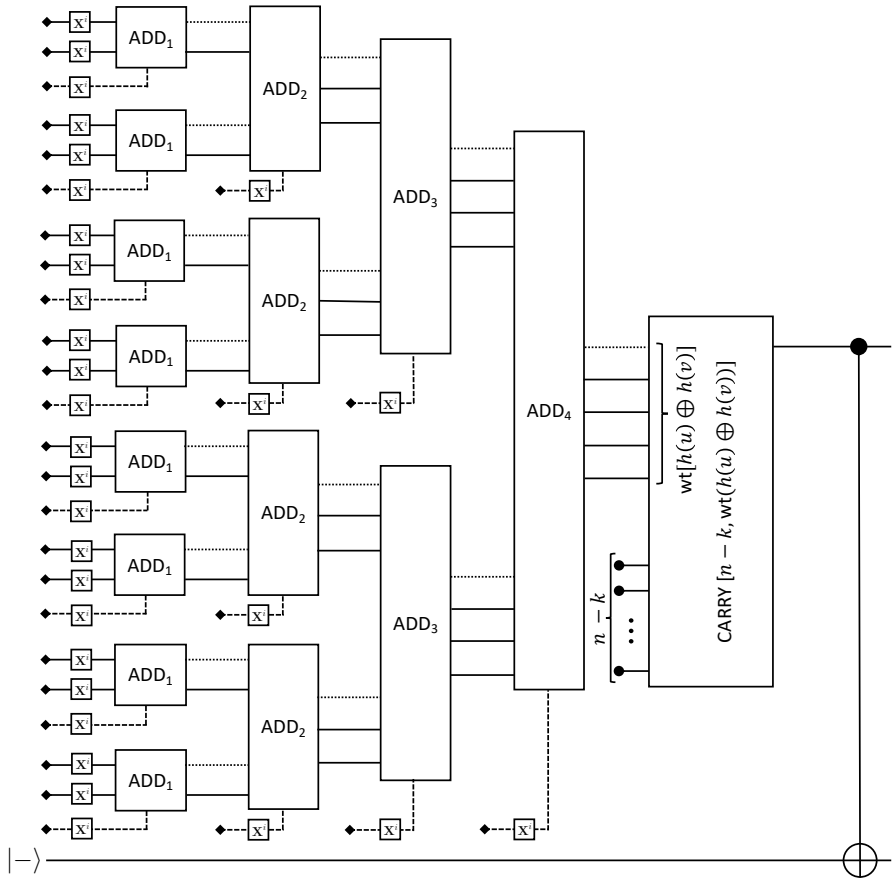



Fig. 1: A quantum circuit for popcount. This circuit computes $h(u) \oplus h(v)$ for a fixed $n$ bit $h(u)$, computes the Hamming weight of $h(u) \oplus h(v)$, and checks whether the Hamming weight is less than or equal to $k$. Here $n = 2^\ell - 1 = 31$. The input qubits are represented as lines ending with a black diamond. The dashed lines represent incoming carry inputs, and the dotted lines represent carry outputs. Not all of the output wires are drawn. For space efficiency, some of the input qubits are fed into the incoming carry qubits of the adders (dashed lines). The $\mathbf{X}^i$ mean that gate $\mathbf{X}$ is applied to input qubit $i$ if bit $i$ of $h(u)$ is 1. The circuit uses a depth $\ell - 1$ binary tree of full bit adders from [16], where $\text{ADD}_i$ denotes an $i$ bit full adder. The output $wt(h(u) \oplus h(v))$ from the tree of adders together with the binary representation of the number $n - k$ are finally fed into the input of the CARRY circuit from [26], which computes the carry bit of $n - k + wt(h(u) \oplus h(v))$ (the carry bit will be 0 if $wt(h(u) \oplus h(v)) \leq k$, and 1 otherwise). The final **CNOT** is for illustration only. In actuality, the carry bit is computed directly into an ancilla that is initialised in the $|-\rangle = (|0\rangle - |1\rangle)/\sqrt{2}$ state, so we can obtain the needed phase kickback. The tree of adders and the initial $\mathbf{X}$ gates, but not the CARRY circuit, are run in reverse to clean up scratch space and return the inputs to their initial state. The uncomputation step is not depicted here.

suppress the ratio of inner products to `popcounts` (see Remark 2). Hence we only need a rough estimate for the cost of an inner product. We assume 32 bits of precision are sufficient. We then assume schoolbook multiplication is used for scalar products, which costs approximately $32^2$ AND instructions. We then assume the cost of a full inner product is approximately $32^2\, d$, i.e. we ignore the cost of the final summation, assuming it is dwarfed by the ANDs.[4]

## 5 The accuracy of `popcount`

Here we give an analysis of the `popcount` technique based on some standard simplifying assumptions. We are particularly interested in the probability that a `popcount` filter identifies a random pair of points as potential neighbours. We are also interested in the probability that a pair of actual neighbours are not identified as potential neighbours, i.e. the false negative rate. Our software computes all of the quantities in this section to high precision.

Let $P_{k,n}(u, v)$ be the probability that $\texttt{popcount}_{k,n}(u, v; h) = 0$ for a uniformly random $h$ (recall $\texttt{popcount}_{k,n}(u, v; h) = 0$ if $u, v$ pass the filter). In other words, let $h = (h_1, \ldots, h_n)$ be a collection of independent random variables that are distributed uniformly on the sphere, and define

$$P_{k,n}(u, v) = 1 - \mathbb{E}\left[\texttt{popcount}_{k,n}(u, v; h)\right].$$

The hyperplane defined by $h_i$ separates $u$ and $v$ with probability $\theta(u, v)/\pi$, and $\texttt{popcount}_{k,n}(u, v; h) = 0$ if no more than $k$ of the hyperplanes separate $u$ and $v$. Hence,

$$P_{k,n}(u, v) = \sum_{i=0}^{k} \binom{n}{i} \cdot \left(\frac{\theta(u, v)}{\pi}\right)^i \cdot \left(1 - \frac{\theta(u, v)}{\pi}\right)^{n-i}.$$

Note that $P_{k,n}(u, v)$ depends only on the angle between $u$ and $v$, so it makes sense to define $P_{k,n}(\theta)$. The main heuristic in our analysis of `popcount` is that $P_{k,n}(u, v)$ is a good approximation to the probability that $\texttt{popcount}_{k,n}(u, v; h) = 0$ for *fixed* $h$ and *varying* $u$ and $v$. Under this assumption, all of the quantities in question can be determined by integrating $P_{k,n}(u, v)$ over different regions of the sphere.

Let $\hat{P}_{k,n}$ denote the event that $\texttt{popcount}_{k,n}(u, v; h) = 0$ for uniformly random $u$, $v$, and $h$. Let $\hat{R}_\theta$ be the event that $\theta(u, v) \leq \theta$. Recall that $\Pr[\hat{R}_\theta] = C_d(\theta)$, and observe that $\Pr[\hat{R}_\theta]$ is a cumulative distribution with associated density

[4] We also tested the effect of assuming 8-bit inner products are sufficient. As expected, this reduces all costs by a factor of two to four and thus does not substantially alter our relative results.

$A_d(\theta) = \frac{\partial}{\partial\theta} C_d(\theta)$. We find, letting $\mathcal{S} = \mathcal{S}^{d-1}$ for some implicit $d$,

$$\begin{aligned}\Pr[\hat{P}_{k,n}] &= \int_{\mathcal{S}} \int_{\mathcal{S}} P_{k,n}(u,v)\ \mathrm{d}\mu(v)\ \mathrm{d}\mu(u) \\ &= \int_{\mathcal{S}} \left( \int_0^{\pi} P_{k,n}(\theta) \cdot A_d(\theta)\ \mathrm{d}\theta \right)\ \mathrm{d}\mu(u) \\ &= \int_0^{\pi} P_{k,n}(\theta) \cdot A_d(\theta)\ \mathrm{d}\theta. \end{aligned} \tag{4}$$

Let $u, v$ such that $\theta(u,v) \leq \varphi$ be neighbours. The false negative rate is $1 - \Pr[\hat{P}_{k,n} \mid \hat{R}_{\varphi}]$. The quantity $\Pr[\hat{P}_{k,n} \wedge \hat{R}_{\varphi}]$ can be calculated by changing the upper limit of integration in Equation 4. It follows that

$$1 - \Pr[\hat{P}_{k,n} \mid \hat{R}_{\varphi}] = 1 - \frac{1}{C_d(\varphi)} \int_0^{\varphi} P_{k,n}(\theta) \cdot A_d(\theta)\ \mathrm{d}\theta. \tag{5}$$

In Section 6 we consider $u$ and $v$ that are uniformly distributed in a cap of angle $\beta < \pi/2$, rather than the uniformly distributed on the sphere. Let $\hat{B}_{w,\beta}$ be the event that $u$ and $v$ are uniformly distributed in a cap of angle $\beta$ about $w$. We have

$$\begin{aligned}\Pr[\hat{B}_{w,\beta}] &= \int_{\mathcal{S}} \int_{\mathcal{S}} \mathbb{1}\left\{ w \in \mathcal{W}^{d-1}(u,\beta,v,\beta) \right\}\ \mathrm{d}\mu(v)\ \mathrm{d}\mu(u) \\ &= \int_0^{2\beta} W_d(\theta,\beta,\beta) \cdot A_d(\theta)\ \mathrm{d}\theta. \end{aligned} \tag{6}$$

In the second line we have used the fact that $\beta < \pi/2$ and $W(\theta, \theta_1, \theta_2)$ is zero when $\theta \geq \theta_1 + \theta_2$. The quantity $\Pr[\hat{B}_{w,\beta} \wedge \hat{R}_{\varphi}]$ can be computed by changing the upper limit of integration in Equation 6 from $2\beta$ to $\min\{2\beta, \varphi\}$. We note that $\hat{B}_{w,\beta}$ has no dependence on $w$ and therefore may also be written $\hat{B}_{\beta}$. The conditional probability that $\texttt{popcount}_{k,n}(u,v;h) = 0$, given that $u, v$ are uniformly distributed in a cap $B_{\beta}$, $\Pr[\hat{P}_{k,n} \mid \hat{B}_{\beta}]$, can be computed using Equation 6 and

$$\Pr[\hat{P}_{k,n} \wedge \hat{B}_{\beta}] = \int_0^{2\beta} P_{k,n}(\theta) \cdot W_d(\theta,\beta,\beta) \cdot A_d(\theta)\ \mathrm{d}\theta. \tag{7}$$

The quantity $\Pr[\hat{P}_{k,n} \wedge \hat{B}_{\beta} \wedge \hat{R}_{\varphi}]$ can be computed by changing the upper limit of integration in Equation 7 from $2\beta$ to $\min\{2\beta, \varphi\}$. The false negative rate for `popcount` when restricted to a cap is $1 - \Pr[\hat{P}_{k,n} \mid \hat{B}_{w,\beta} \wedge \hat{R}_{\varphi}]$.

## 6 Tuning `popcount` for NNS

We now use the circuit sizes from Section 4 and the probabilities from Section 5 to optimise `popcount` for use in NNS algorithms. Our analysis is with respect to points sampled independently from the uniform distribution on the sphere.

We further restrict our attention to *list-size preserving* parameterisations, which take an input list of size $N$ and return an output list of (expected) size $N$.

We use the notation for events introduced in Section 5. In particular, we write $\hat{R}_\theta$ for the event that a uniformly random pair of vectors are neighbours, i.e. that they lie at angle less than or equal to $\theta$ of one another; $\hat{P}_{k,n}$ for the event that `popcount` identifies a uniformly random pair of vectors as potential neighbours; $\hat{B}_\beta$ for the event that a uniformly random pair of vectors lie in a uniformly random cap of angle $\beta$; and $\hat{B}_{w,\beta}$ for the same event except we highlight the cap is centred on $w$. Throughout this section we use $\texttt{popcount}_{k,n}(u,\cdot)$, for various fixed $u$, as a filter for the search predicate $\theta(u,\cdot) \leq \theta$. We write $\eta(k,n)$ for the false negative rate of `popcount`. We assume that $\theta(u,v) \leq \theta$ is computed using an inner product test. Throughout this section, $c_1$ represents the instruction cost of the inner product test from Section 4.3, $c_2(k,n)$ the instruction cost of `popcount` from Section 4.2, $q_1$ the quantum cost of the reflection $\mathbf{R}_{f\cap g}$, and $q_2(k,n)$ the quantum cost of $\mathbf{G}(g)$ from Section 4.1. We note that $c_1, q_1$ have a dependence on $d$ that we suppress. We write $q_0(m)$ for the number of $\mathbf{G}(g)$ iterations that are applied during a quantum search on a set of size $m$.

Our goal is to minimise the cost of list-size preserving NNS algorithms as a function of the input list size, the `popcount` parameters $k$ and $n$, and the other NNS parameters. In a list of $N$ points there are $\binom{N}{2}$ ordered pairs. We expect $\binom{N}{2} \cdot \Pr[\hat{R}_\theta] = \binom{N}{2} \cdot C_d(\theta)$ of these to be neighbours, and we expect a $1-\eta(k,n)$ fraction of neighbours to be detected by `popcount`. List-size preserving parmaterisations that use a `popcount` filter must therefore take an input list of size at least

$$\ell(k,n) = \frac{2}{1-\eta(k,n)} \cdot \frac{1}{C_d(\theta)}. \tag{8}$$

The optimised costs reported in Figure 2 typically use popcount parameters for which $\ell(k,n) \in (2/C_d(\pi/3), 4/C_d(\pi/3))$. Here we assume that list-size preserving parameterisations take $N = \ell(k,n)$. Note that $\eta(k,n) = 1 - \Pr[\hat{P}_{k,n} \mid \hat{R}_\theta]$ when the search is over a set of points uniformly distributed on the sphere, and $\eta(k,n) = 1 - \Pr[\hat{P}_{k,n} \mid \hat{R}_\theta \wedge \hat{B}_\beta]$ when the search is over a set of points uniformly distributed in a cap of angle $\beta$ (left implicit).

In each of the quantum analyses, we apply Proposition 1 with $\gamma = 1$, $P = |g|$ and $Q = 1$ to estimate $q_0(m)$. We assume that filtered quantum search succeeds with probability 1 instead of probability at least $1/14$, as guaranteed by Proposition 1. In practice, one will not know $|g|$ and one will therefore take $\gamma > 1$. Our use of $\gamma = 1$ is a systematic underestimate of the true cost of the search. There may be searches where our lower bound of $Q = 1$ on $|f \cap g|$ is too pessimistic. However, the probability of success in filtered quantum search decreases quadratically with $Q/|f \cap g|$ if $Q > |f \cap g|$. In Sections 6.1 and 6.3 we expect $|f \cap g| \approx 2$ so the effect of taking $Q = 1$ is negligible. In Section 6.2, where $Q$ may be larger, an optimistic analysis using the expected value of $Q$ makes negligible savings in dimension 512 and small savings in dimension 1024. This analysis does not decrement $Q$ when a neighbour is found in, then removed from, a search space and ignores the quadratic decrease in success probability.

### 6.1 AllPairSearch

As a warmup, we optimise AllPairSearch. Asymptotically its complexity is $2^{(0.415\ldots+o(1))d}$ classically and $2^{(0.311\ldots+o(1))d}$ quantumly. We describe implementations of Line 5 of Algorithm 1 based on filtered search and filtered quantum search, and optimise `popcount` relative to these implementations.

**Filtered search.** Suppose that Line 5 applies $\texttt{popcount}_{k,n}(v_i, \cdot)$ to each of $v_{i+1}$ through $v_N$ and then applies an inner product test to each vector that passes. With an input list of size $N = \ell(k, n)$, we expect this implementation to test all $\binom{N}{2}$ pairs before finding $N$ neighbouring pairs. Moreover, we expect the `popcount` filter to identify $\binom{N}{2} \cdot \Pr[\hat{P}_{k,n}]$ potential neighbours, and to perform an equal number of inner product tests. The optimal parameters are obtained by minimising

$$\left(c_1 \cdot \Pr[\hat{P}_{k,n}] + c_2(k, n)\right) \cdot \binom{\ell(k,n)}{2}. \tag{9}$$

**Filtered quantum search.** Suppose that Line 5 is implemented using the search routine Algorithm 2. Specifically, we take the predicate $f$ to be $\theta(v_i, \cdot) \leq \theta$ with domain $L_i$. We take the filter $g$ to be $\texttt{popcount}_{k,n}(v_i, \cdot)$. Each call to the search routine returns at most one neighbour of $v_i$. To find all detectable neighbours of $v_i$ in $L_i$ we must repeat the search $|f \cap g|$ times. This is expected to be $|L_i| \cdot \Pr[\hat{P}_{k,n} \wedge \hat{R}_\theta]$. Known neighbours of $v_i$ can be removed from $L_i$ to avoid a coupon collector scenario. We consider an implementation in which searches are repeated until a search fails to find a neighbour of $v_i$.

We expect to call the search subroutine $|L_i| \cdot \Pr[\hat{P}_{k,n} \wedge \hat{R}_\theta] + 1$ times in iteration $i$. Proposition 1 with $P = |L_i| \cdot \Pr[\hat{P}_{k,n}]$, $Q = 1$, and $\gamma = 1$ gives $q_0(|L_i|) = \frac{1}{2}\sqrt{|L_i|}$ iterations of $\mathbf{G}(g)$. As $i$ ranges from 1 to $N-1$ the quantity $|L_i|$ takes each value in $\{1, \ldots, N-1\}$. Our proposed implementation therefore performs an expected

$$\sum_{j=1}^{N-1} \frac{1}{2}\sqrt{j}\left(j \cdot \Pr[\hat{P}_{k,n} \wedge \hat{R}_\theta] + 1\right)$$
$$= \Pr[\hat{P}_{k,n} \wedge \hat{R}_\theta]\left(\frac{1}{5}N^{5/2} + \frac{1}{4}N^{3/2}\right) + \frac{1}{3}N^{3/2} + O(\sqrt{N}) \tag{10}$$

applications of $\mathbf{G}(g)$; the expansion is obtained by the Euler–Maclaurin formula. When $N = \ell(k, n)$ we expect $N \cdot \Pr[\hat{P}_{k,n} \wedge \hat{R}_\theta] = 2 + O(1/N)$. The right hand side of Equation 10 is then $\frac{11}{15}N^{3/2} + O(\sqrt{N})$.

Proposition 1 also provides an estimate for the rate at which reflections about the true positives, $\mathbf{R}_{f \cap g}$ are performed. With $P$ and $Q$ as above, we find that $\mathbf{R}_{f \cap g}$ is performed at roughly $p(k, n) = \sqrt{\Pr[\hat{P}_{k,n}]}$ the rate of calls to $\mathbf{G}(g)$. The

optimal popcount parameters (up to some small error due to the $O(\sqrt{N})$ term in Equation 10) are obtained by minimising the total cost

$$\frac{11}{15}\left(q_1 p(k,n) + q_2(k,n)\right) \cdot \ell(k,n)^{3/2}. \tag{11}$$

### 6.2 RandomBucketSearch

One can improve AllPairSearch by *bucketing* the search space such that vectors in the same bucket are more likely to be neighbours [33]. For example, one could pick a hemisphere $H$ and divide the list into $L_1 = L \cap H$ and $L_2 = L \backslash L_1$. These lists would be approximately half the size of the original and the combined cost of AllPairSearch within $L_1$ and then within $L_2$ would be half the cost of an AllPairSearch within $L$. However, this strategy would fail to detect the expected $\theta/\pi$ fraction of neighbours that lie in opposite hemispheres.

Becker, Gama, and Joux [9] present a very efficient generalisation of this strategy. They propose bucketing the input list into subsets of the form $\{v \in L : \texttt{popcount}_{k,n}(0, v; h) = 0\}$ with varying choices of $h$. This bucketing strategy is applied recursively until the buckets are of a minimum size. Neighbouring pairs are then found by an AllPairSearch.

A variant of the Becker–Gama–Joux algorithm that uses buckets of the form $L \cap \mathcal{C}^{d-1}(f, \theta_1)$, with randomly chosen $f$ and fixed $\theta_1$, was proposed and implemented in [2]. This variant is sometimes called bgj1. Here we call it RandomBucketSearch. This algorithm has asymptotic complexity $2^{(0.349\ldots+o(1))d}$ classically [2] and $2^{(0.301\ldots+o(1))d}$ quantumly.[5] This is worse than the Becker–Gama–Joux algorithm, but RandomBucketSearch is conceptually simple and still provides an enormous improvement over AllPairSearch. Pseudocode is presented in Algorithm 3.

**Description of Algorithm 3.** The algorithm takes as input a list of $N$ points uniformly distributed on the sphere. A random bucket centre $f$ is drawn uniformly from $\mathcal{S}^{d-1}$ in each of the $t$ iterations of the outer loop. The choice of $f$ defines a bucket in Line 5, $L_f = L \cap \mathcal{C}^{d-1}(f, \theta_1)$, which is of expected size $N \cdot C_d(\theta_1)$. For each $v_j \in L_f$, the inner loop searches a set $L_{f,j} \subset L_f$ for neighbours of $v_j$. The quantity $|L_{f,j}|$ takes each value in $\{1, \ldots, |L_f| - 1\}$ as $v_j$ ranges over $L_f$. The inner loop is identical to the loop in AllPairSearch apart from indexing and the fact that elements of $L_f$ are known to be in the cap $\mathcal{C}^{d-1}(f, \theta_1)$.

A bucket $L_f$ is expected to contain $\binom{N}{2} \cdot \Pr[\hat{R}_\theta \wedge \hat{B}_{f,\theta_1}]$ neighbouring pairs. Only a $1 - \eta(k,n)$ fraction of these are expected to be identified by the popcount filter. When $\theta_1 > \theta$ it is reasonable to assume that $\Pr[\hat{R}_\theta \wedge \hat{B}_{f,\theta_1}] \approx C_d(\theta) \cdot W_d(\theta, \theta_1, \theta_1)$. We use this approximation. The expected number of neighbouring

[5] The asymptotic quantum complexity is calculated, similarly to the classical complexity [2], using the asymptotic value of $W_d(\theta, \theta_1, \theta_1)$ given in [8]. Let $N = 1/C_d(\pi/3)$ and $t(\theta_1) = 1/W_d(\pi/3, \theta_1, \theta_1)$. The exponent $0.3013\ldots$ is obtained by minimising $t(\theta_1)\left(N + (NC_d(\theta_1))^{3/2}\right)$ with respect to $\theta_1$.

**Algorithm 3** RandomBucketSearch

**Input:** A list $L = (v_1, v_2, \dots v_N) \subset \mathcal{S}^{d-1}$ of $N$ points. Parameters $\theta, \theta_1 \in (0, \pi/2)$ and $t \in \mathbb{Z}_+$.
**Output:** A list of pairs $(u, v) \in L \times L$ with $\theta(u, v) \leq \theta$.

1: **function** RandomBucketSearch($L; \theta, \theta_1, t$)
2: $L' \leftarrow \emptyset$
3: **for** $1 \leq i \leq t$ **do**
4: Sample $f$ uniformly on $\mathcal{S}^{d-1}$
5: $L_f \leftarrow L \cap \mathcal{C}^{d-1}(f, \theta_1)$
6: **for** $j$ such that $v_j \in L_f$ **do**
7: $L_{f,j} \leftarrow \{v_k \in L_f : j < k \leq N\}$
8: Search $L_{f,j}$ for any number of $u$ that satisfy $\theta(v_j, u) \leq \theta$
9: For each such $u$ found, add $(v_j, u)$ to $L'$.
10: If $|L'| \geq N$, **return** $L'$.
11: **return** $L'$

pairs in $L_f$ that are detected by the `popcount` filter is therefore approximately $\binom{N}{2} \cdot (1 - \eta(k,n)) \cdot C_d(\theta) \cdot W_d(\theta, \theta_1, \theta_1)$. When $N = \ell(k,n)$ this is $N \cdot W_d(\theta, \theta_1, \theta_1)$. If all detectable neighbours are found by the search routine then the algorithm is list-size preserving when $N = \ell(k,n)$ and $t = 1/W_d(\theta, \theta_1, \theta_1)$.

We can now derive optimal `popcount` parameters for various implementations of Line 8.

**Filtered search.** Suppose that Line 8 of Algorithm 3 applies $\texttt{popcount}_{k,n}(v_j, \cdot)$ to each element of $L_{f,j}$ and then applies an inner product test to each vector that passes. This implementation applies `popcount` tests to all $\binom{|L_f|}{2} \approx \binom{N \cdot C_d(\theta_1)}{2}$ pairs of elements in $L_f$ and finds all of the neighbouring pairs that pass. In the process it applies inner product tests to a $p(\theta_1, k, n) = \Pr[\hat{P}_{k,n} \mid \hat{B}_{f,\theta_1}]$ fraction of pairs. The cost of populating buckets in one iteration of Line 5 is $c_1 \cdot \ell(k,n)$. The cost of all searches on Line 8 is $(c_1 \cdot p(\theta_1, k, n) + c_2(k,n)) \cdot \binom{NC_d(\theta_1)}{2}$. With the list-size preserving parameters $N$ and $t$ given above, the optimal $\theta_1$, $k$, and $n$ can be obtained by minimising the total cost

$$\frac{c_1 \cdot \ell(k,n) + (c_1 \cdot p(\theta_1, k, n) + c_2(k,n)) \cdot \binom{\ell(k,n) \cdot C_d(\theta_1)}{2}}{W_d(\theta, \theta_1, \theta_1)}. \tag{12}$$

**Filtered quantum search.** Suppose that Line 8 is implemented using the search routine Algorithm 2. We take the predicate $f$ to be $\theta(v_j, \cdot) \leq \theta$ with domain $L_{f,j}$. We take the filter $g$ to be $\texttt{popcount}_{k,n}(v_j, \cdot)$. Each call to the search routine returns at most one neighbour of $v_j$. To find all detectable neighbours of $v_j$ in $L_{f,j}$ we must repeat the search several times. Known neighbours of $v_j$ can be removed from $L_{f,j}$ to avoid a coupon collector scenario. Proposition 1 with $P = |L_{f,j}| \cdot \Pr[\hat{P}_{k,n} \mid \hat{B}_{f,\theta_1}]$, $Q = 1$, and $\gamma = 1$ gives us that the number of $\mathbf{G}(g)$ iterations in a search on a set of size $|L_{f,j}|$ is $q_0(|L_{f,j}|) = \frac{1}{2}\sqrt{|L_{f,j}|}$.

We consider an implementation of Line 8 in which searches are repeated until a search fails to find a neighbour of $v_j$. With $N = \ell(k,n)$, the set $L_f$ is of expected size $\ell(k,n) \cdot C_d(\theta_1)$ and contains an expected $\ell(k,n) \cdot W_d(\theta,\theta_1,\theta_1)$ neighbouring pairs detectable by `popcount`. The set $L_{f,j}$ is expected to contain a proportional fraction of these pairs. As such, we expect to call the search subroutine $|L_{f,j}| \cdot r(\theta_1,k,n) + 1$ times in iteration $j$ where

$$r(\theta_1,k,n) = \frac{N \cdot W_d(\theta,\theta_1,\theta_1)}{\binom{|L_f|}{2}} \approx \frac{2\,W_d(\theta,\theta_1,\theta_1)}{\ell(k,n) \cdot C_d(\theta_1)^2}.$$

The inner loop makes an expected

$$\sum_{j=1}^{|L_f|-1} \frac{1}{2}\sqrt{j}\,(j \cdot r(\theta_1,k,n) + 1)$$

applications of $\mathbf{G}(g)$. This admits an asymptotic expansion similar to that of Equation 10. If we assume that $|L_f|$ takes its expected value of $\ell(k,n) \cdot C_d(\theta_1)$, then the inner loop makes

$$q_3(\theta_1,k,n) \cdot (\ell(k,n) \cdot C_d(\theta_1))^{3/2}$$

applications of $\mathbf{G}(g)$, where

$$q_3(\theta_1,k,n) = \frac{2\,W_d(\theta,\theta_1,\theta_1)}{5\,C_d(\theta_1)} + \frac{1}{3}.$$

Proposition 1 also provides an estimate for the rate at which reflections about the true positives, $\mathbf{R}_{f\cap g}$ are performed. With $P$ and $Q$ as above, we find that $\mathbf{R}_{f\cap g}$ is applied at roughly $p(\theta_1,k,n) = \sqrt{\Pr[\hat{P}_{k,n} \mid \hat{B}_{f,\theta_1}]}$ the rate of $\mathbf{G}(g)$ iterations. The total cost of searching for neighbouring pairs in $L_f$ is therefore

$$s(\theta_1,k,n) = (q_1 \cdot p(\theta_1,k,n) + q_2(k,n)) \cdot q_3(\theta_1,k,n) \cdot \left(\ell(k,n) \cdot C_d(\theta_1)\right)^{3/2}. \quad (13)$$

Populating $L_f$ has a cost of $c_1 \cdot \ell(k,n)$. With the list-size preserving $t$ given above, the optimal parameters $\theta_1$, $k$, and $n$ can be obtained by minimising the total cost

$$\frac{c_1 \cdot \ell(k,n) + s(\theta_1,k,n)}{W_d(\theta,\theta_1,\theta_1)}. \quad (14)$$

### 6.3 ListDecodingSearch

The optimal choice of $\theta_1$ in RandomBucketSearch balances the cost of $N \cdot t$ cap membership tests against the cost of all calls to the search subroutine. It can be seen that reducing the cost of populating the buckets would allow us to choose a smaller $\theta_1$, which would reduce the cost of searching within each bucket.

Algorithm 4, ListDecodingSearch, is due to Becker, Ducas, Gama, and Laarhoven [8]. Its complexity is $2^{(0.292\ldots+o(1))d}$ classically and $2^{(0.265\ldots+o(1))d}$ quantumly [34, 35]. Like RandomBucketSearch, it computes a large number of list-cap intersections. However, these list-cap intersections involve a structured list—the list-cap intersections in RandomBucketSearch involve the inherently unstructured input list.

**Algorithm 4** ListDecodingSearch

**Input:** A list $L = (v_1, v_2, \ldots v_N) \subset \mathcal{S}^{d-1}$ of $N$. Parameters $\theta, \theta_1, \theta_2 \in (0, \pi/2)$ and $t \in \mathbb{Z}_+$.

**Output:** A list of pairs $(u, v) \in L \times L$ with $\theta(u, v) \leq \theta$.

1: **function** ListDecodingSearch($L; \theta, \theta_1, \theta_2, t$)
2: Sample a random product code $F$ of size $t$
3: Initialise an empty list $L_f$ for each $f \in F$
4: **for** $1 \leq i \leq N$ **do**
5: $F_i \leftarrow F \cap \mathcal{C}^{d-1}(v_i, \theta_2)$
6: Add $v_i$ to $L_f$ for each $f$ in $F_i$
7: **for** $1 \leq j < N$ **do**
8: $F_j \leftarrow F \cap \mathcal{C}^{d-1}(v_j, \theta_1)$
9: **for** $f \in F_j$ **do**
10: $L_{f,j} \leftarrow \{v_k \in L_f : j < k \leq N\}$
11: $L_{F,j} \leftarrow \coprod_{f \in F_j} L_{f,j}$ (disjoint union)
12: Search $L_{F,j}$ for any number of $u$ that satisfy $\theta(v_j, u) \leq \theta$
13: For each such $u$ found, add $(v_j, u)$ to $L'$.
14: If $|L'| \geq N$, **return** $L'$.
15: **return** $L'$

**Description of Algorithm 4.** The algorithm first samples a $t$ point *random product code* $F$. See [8] for background on random product codes. In our analysis, we treat $F$ as a list of uniformly random points on $\mathcal{S}^{d-1}$. A formal statement is given as [8, Theorem 5.1], showing that such a heuristic is essentially true, up to a subexponential loss on the probability of finding the intend pairs.

The first loop populates $t$ buckets that have as centres the points $f$ of $F$. Bucket $L_f$ stores elements of $L$ that lie in the cap of angle $\theta_2$ about $f$. Each bucket is of expected size $N \cdot C_d(\theta_2)$.

The second loop iterates over $v_j \in L$ and searches for neighbours of $v_j$ in the disjoint union of buckets with centres within an angle $\theta_1$ of $v_j$. The set $F_j$ constructed on Line 8 contains an expected $t \cdot C_d(\theta_1)$ bucket centres. The disjoint union of certain elements from the corresponding buckets, denoted $L_{F,j}$, is of expected size $(N-j) \cdot C_d(\theta_2) \cdot t \cdot C_d(\theta_1)$. We note that by simplifying and assuming the expected size of $L_{F,j}$ is $N \cdot C_d(\theta_2) \cdot t \cdot C_d(\theta_1)$ the costs given below are never wrong by more than a factor of two.

Suppose that $w$ is a neighbour of $v_j$, so $\theta(v_j, w) \leq \theta$. The measure of the wedge formed by a cap of angle $\theta_1$ about $v_j$ and a cap of angle $\theta_2$ about $w$ is at least $W_d(\theta, \theta_1, \theta_2)$. Assuming that the points of a random product code are indistinguishable from points sampled uniformly on the sphere, the probability that some $f \in F_j$ contains $w$ is at least $t \cdot W_d(\theta, \theta_1, \theta_2)$.

The second loop is executed $N$ times. Iteration $j$ searches $L_{F,j}$ for neighbours of $v_j$. With $N = \ell(k, n)$ there are expected to be $N$ detectable neighbouring pairs in $L$. With $t = 1/W_d(\theta, \theta_1, \theta_2)$ we expect that each neighbouring pair is of the form $(v_j, w)$ with $w \in L_{F,j}$.

The angles $\theta_1, \theta_2$ relate to the spherical cap parameters $\alpha, \beta$ respectively in [8], and are such that $\theta_1 \geq \theta_2$. Optimal time complexity is achieved when $\theta_1 = \theta_2$.

We have omitted the list decoding mechanism by which list-cap intersections are computed. In our analysis we assume that the cost of a list-cap intersection such as $F_i = F \cap \mathcal{C}^{d-1}(v_i, \theta_2)$ is proportional to $|F_i|$, but independent of $|F|$, i.e. we are in the "efficient list-decodability regime" of [8, Section 5.1] and may take their parameter $m = \log d$. In particular, we assume that in the cost of $O(\log(d) \cdot |F_i|)$ inner products and $|F|^{O(1/\log(d))}$ other operations, as stated in [8, Lemma 5.1], the first cost dominates. In [8] these costs relate to $O(m \cdot M \cdot \mathcal{C}_n(\alpha))$ and $O(nB + mB \log B)$ respectively. We therefore assume the cost of forming $F_i = F \cap \mathcal{C}^{d-1}(v_i, \theta_2)$ is $\log(d) \cdot |F_i|$ inner product tests.

**Filtered search.** Suppose that the implementation of Line 12 of Algorithm 4 applies $\texttt{popcount}_{k,n}(v_j, \cdot)$ to each element of $L_{F,j}$ and then applies an inner product test to each vector that passes. This implementation applies `popcount` tests to all $N \cdot C_d(\theta_2) \cdot t \cdot C_d(\theta_1)$ elements of $L_{F,j}$ and finds all of the neighbours of $v_j$ that pass. Note that $w \in L_{F,j}$ implies that there exists some $f \in F$ such that both $v_j$ and $w$ lie in a cap of angle $\theta_1$ around $f$. Inner product tests are applied to a $p(\theta_1, k, n) \geq \Pr[\hat{P}_{k,n} \mid \hat{B}_{f,\theta_1}]$ fraction of all pairs.[6]

The cost of preparing all $t$ buckets in the first loop is $c_1 \cdot N \cdot t \cdot C_d(\theta_2)$. The cost of constructing the search spaces in the second loop is $c_1 \cdot N \cdot t \cdot C_d(\theta_1)$. Each search has a cost of $|L_{F,j}|$ `popcount` tests and $|L_{F,j}| \cdot p(\theta_1, k, n)$ inner product tests. With the list-size preserving parameterisation given above, the optimal $\theta_1$, $\theta_2$, $k$, and $n$ can be obtained by minimising the total cost

$$\frac{\ell(k,n)}{W_d(\theta,\theta_1,\theta_2)}\Big(c_1 \cdot C_d(\theta_1) + c_1 \cdot C_d(\theta_2) + \big(c_1 \cdot p(\theta_1,k,n) + c_2(k,n)\big) \cdot \ell(k,n) \cdot C_d(\theta_1) \cdot C_d(\theta_2)\Big). \quad (15)$$

**Filtered quantum search.** Suppose that Line 12 is implemented using Algorithm 2. We take the predicate $f$ to be $\theta(v_j, \cdot) \leq \theta$ with domain $L_{F,j}$. We take the filter $g$ to be $\texttt{popcount}_{k,n}(v_j, \cdot)$. Each call to the search routine returns at most one neighbour of $v_j$. Known neighbours of $v_j$ can be removed from $L_{F,j}$ to

[6] The inequality is because $v_j$ and $w$ may be contained in multiple buckets, $L_{f,j}$.

avoid a coupon collector scenario. Proposition 1 with $P = |L_{F,j}| \cdot \Pr[\hat{P}_{k,n} \mid \hat{B}_{f,\theta_2}]$, $Q = 1$, and $\gamma = 1$ gives us that the number of $\mathbf{G}(g)$ iterations in a search on a set of size $|L_{F,j}|$ is $q_0(|L_{F,j}|) \approx \frac{1}{2}\sqrt{|L_{F,j}|}$.

Assuming that computing $F_j = F \cap C(v_j, \theta_1)$ has a cost of $c_1 |F_j|$, the $N$ iterations of Lines 5 and 8 have a total cost of

$$c_1 \cdot N \cdot t \cdot (C_d(\theta_1) + C_d(\theta_2)) \tag{16}$$

Each search applies an expected

$$q_0(|L_{F,j}|) \approx \frac{1}{2}\sqrt{N \cdot C_d(\theta_1) \cdot t \cdot C_d(\theta_2)}$$

applications of $\mathbf{G}(g)$. Reflections about the true positives, $\mathbf{R}_{f \cap g}$, are performed at roughly $p(\theta_1, k, n) = \sqrt{\Pr[\hat{P}_{k,n} \mid B_{f,\theta_1}]}$ the rate of $\mathbf{G}(g)$ iterations. We consider an implementation of Line 8 in which searches are repeated until a search fails to find a neighbour of $v_j$. With the list-size preserving parameters given above, we expect to perform two filtered quantum searches per iteration of the second loop. The optimal parameters can be obtained by minimising the total cost

$$\ell(k,n)\left(c_1 \frac{C_d(\theta_1) + C_d(\theta_2)}{W_d(\theta, \theta_1, \theta_2)} + (q_1 p(\theta_1, k, n) + q_2(k,n))\sqrt{\frac{\ell(k,n) C_d(\theta_1) C_d(\theta_2)}{W_d(\theta, \theta_1, \theta_2)}}\right).$$

## 7 Cost estimates

Our software numerically optimises the cost functions in Sections 6.1, 6.2 and 6.3 with respect to several classical and quantum cost metrics. The classical cost metrics that we consider are: *c (unit cost)*, which assigns unit cost to `popcount`; *c (RAM)*, which uses the classical circuits of Section 4. The quantum cost metrics that we consider are: *q (unit cost)*, which assigns unit cost to a Grover iteration; *q (depth-width)*, which assigns unit cost to every gate (including the identity) in the quantum circuits of Section 4; *q (gates)*, which assigns unit cost only to the non-identity gates; *q (T count)*, which assigns unit cost only to $T$ gates; and *q (GE19)*, which is described in Section 7.1.

We stress that our software, and Figure 2, give *estimates* for the cost of each algorithm. These estimates are neither upper bounds nor lower bounds. As we mention above, we have systematically omitted and underestimated some costs. For instance, we have omitted the list decoding mechanism in our costing of Algorithm 4. We have approximated other costs. For instance, the cost that we assign to an inner product in Section 4.3. We have also not explored the entire optimisation space. We only consider values of the popcount parameter $n$ that are one less than a power of two. Moreover, following the discussion in Section 2.4, we set $k = \lfloor n/3 \rfloor$.

While we have omitted and approximated some costs, we have tried to ensure that these omissions and approximations will ultimately lead our software to *underestimate* of the total cost of the algorithm. For instance, if our inner product

cost is accurate, our optimisation procedure ensures that we satisfy Remark 2 and can ignore costs relating to $\mathbf{R}_{f \cap g}$.

Our results are presented in Figure 2. We also plot the leading term of the asymptotic complexity of the respective algorithms as these are routinely referred to in the literature. The source code, and raw data for all considered cost metrics, is available at `https://github.com/jschanck/eprint-2019-1161`.

### 7.1 Barriers to a quantum advantage

As expected, our results in Figure 2 indicate that quantum search provides a substantial savings over classical search asymptotically. Our plots fully contain the range of costs from $2^{128}$ to $2^{256}$ that are commonly thought to be cryptanalytically interesting. Modest cost improvements are attained in this range.

The range of parameters in which a sieve could conceivably be run, however, is much narrower. If one assumes a memory density of one petabyte per gram ($2^{53}$ bits per gram), a $2^{140}$ bit memory would have a mass comparable with that of the Moon. Supposing that a 2-sieve stores $1/C_d(\pi/3)$ vectors, and that each vector is $\log_2(d)$ bits, an adversary with a $2^{140}$ bit memory could only run a sieve in dimension 608 or lower. The potential cost improvement in dimension 608 is smaller than the potential cost improvement in, say, dimension 1000. The potential cost improvement that can be actualised is likely smaller still.

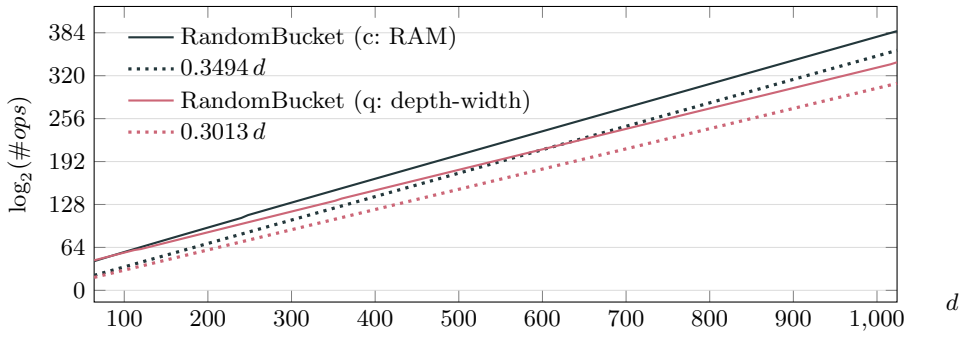
We expect that our cost estimates are underestimates. However, the quantum advantage could grow, shrink, or even be eliminated if our underestimates do not affect quantum and classical costs equally. In this section, we list several reasons to think that the advantage might shrink or disappear.

**Error correction overhead.** By using the depth-width metric for quantum circuits, we assume that dispatching a logical gate to a logical qubit costs one RAM instruction. In practice, however, the cost depends on the error correcting code that is used for logical qubits. This cost may be significant.

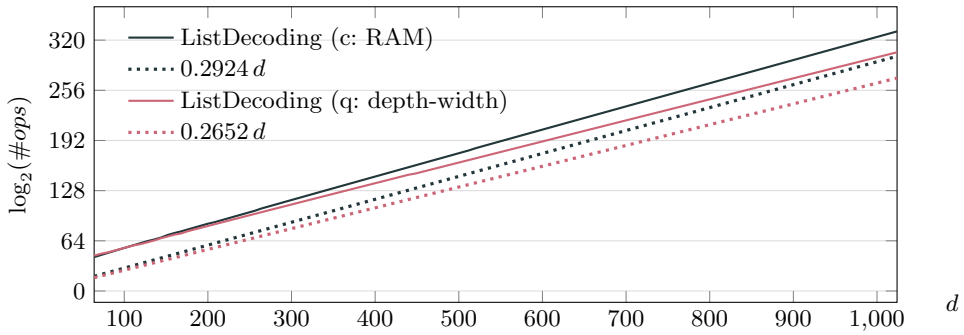
Gidney and Ekerå have estimated the resources required to factor a 2048 bit RSA modulus using Shor's algorithm on a surface code based quantum computer [20]. Under a plausible assumption on the physical qubit error rate, they calculate that a factoring circuit with $2^{12.6}$ logical qubits and depth $2^{31}$ requires a distance $\delta = 27$ surface code. Each logical qubit is encoded in $2\,\delta^2 = 1458$ physical qubits, and the error tracking routine applies at least $\delta^2 = 729$ bit instructions, per logical qubit per layer of logical circuit depth, to read its input.

In general, a circuit of depth $D$ and width $W$ requires a distance $\delta = \Theta(\log(DW))$ surface code. To perform a single logical gate, classical control hardware dispatches several instructions to each of the $\Theta(\log^2(DW))$ physical qubits. The classical control hardware also performs a non-trivial error tracking routine between logical gates, which takes measurement results from half of the

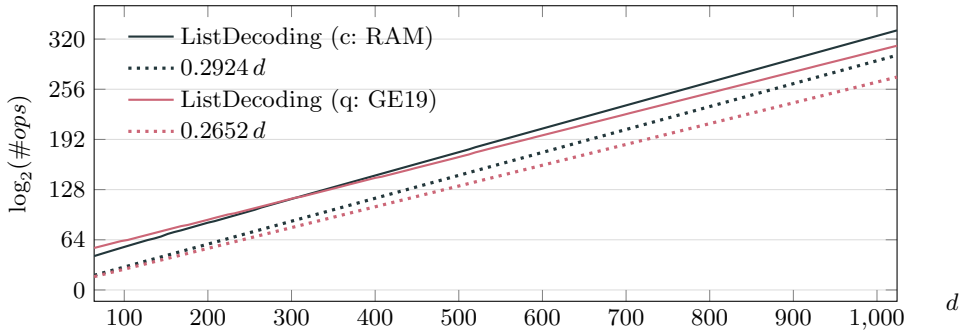
Fig. 2: Quantum ("q") and classical ("c") resource estimates for NNS search.




RandomBucketSearch. Comparing c: (RAM) with q: (depth-width), and the leading terms of the asymptotic complexities.




ListDecodingSearch. Comparing c: (RAM) with q: (depth-width), and the leading terms of the asymptotic complexities.




ListDecodingSearch. Comparing c: (RAM) with q: (GE19), and the leading terms of the asymptotic complexities.

physical qubits as input.[7] Consequently, the cost of surface code computation grows like $\Omega(DW \log^2(DW))$.

We have adapted scripts provided by Gidney and Ekerå to estimate $\delta$ for our circuits. The last plot of Figure 2 shows the cost of ListDecodingSearch when every logical gate (including the identity) is assigned a cost of $\delta^2$. For ListDecodingSearch the cost in the Gidney–Ekerå metric grows from $2^{128}$ to $2^{256}$ between dimensions 352 and 824, and we calculate a $2^{128}$ bit memory is sufficient to run in dimension 544. We find that the advantage of quantum search over classical search is a factor of $2^{1.8}$ in dimension 352, a factor of $2^{7.1}$ in dimension 544, and a factor of $2^{14.4}$ in dimension 824. Compare this with the naïve estimate for the advantage, $2^{0.292d-0.265d}$, which is a factor of $2^{9.5}$ in dimension 352, a factor of $2^{14.7}$ in dimension 544, and a factor of $2^{22.5}$ in dimension 824.

One should also note that error correction for the surface code sets a natural clock speed, which Gidney and Ekerå estimate at one cycle per microsecond. Gidney and Ekerå estimate that their factoring circuit, the cost of which is dominated by a single modular exponentiation, would take 7.44 hours to run. This additional overhead in terms of time is not refelected in the instruction count.

On the positive side, the cost estimate used in Figure 2 is specific to the surface code architecture. Significant improvements may be possible. Gottesman has shown that an overhead of $\Theta(1)$ physical qubits per logical qubit is theoretically possible [22]. Whether this technique offers lower overhead than the surface code in practice is yet to be seen.

**Dependence on qRAM.** Quantum accessible classical memories are used in many quantum algorithms. For example, they are used in black box search algorithms [25], in collision finding algorithms [14], and in some algorithms for the the dihedral hidden subgroup problem [32]. The use of qRAM is not without controversy [11, 24]. Previous work on quantum lattice sieve algorithms [34, 35] has noted that constructing practical qRAM seems challenging.

Morally, looking up an $\ell$ bit value in a table with $2^n$ entries should have a cost that grows at least with $n + \ell$. Recent results [5, 6, 38] indicate that realistic implementations of qRAM have costs that grow much more quickly than this. When ancillary qubits are kept to a minimum, the best known Clifford+T implementation of a qRAM has a $\mathbf{T}$ count of $4 \cdot (2^n - 1)$ [6]. While it is conceivable that a qRAM could be constructed at lower cost on a different architecture, as has been suggested in [21], a unit cost qRAM gate should be seen as a powerful, and potentially unrealistic, resource.

One can argue that classical RAMs also have a large cost. This is not to say that classical and quantum RAMs have the same cost. A qRAM can be used to construct an arbitrary superposition over the elements of a memory. This process relies on quantum interference and necessarily takes as long as a worst case memory access time. This is in contrast with classical RAM, where

[7] For a thorough introduction to how logical gates are performed on the surface code see [19], and for more advanced techniques see e.g. [27].

careful programming and attention to a computer's caches can mask the fact that accessing an $N$ bit memory laid out in a 3-dimensional space necessarily takes $\Omega(N^{1/3})$ time.

If the cost of a qRAM gate is equivalent to $\Theta(N^{1/3})$ Clifford+T gates, then the asymptotic cost of quantum AllPair search is $2^{(0.380\ldots+o(1))d}$, the asymptotic cost of quantum RandomBucket search is $2^{(0.336\ldots+o(1))d}$, and the asymptotic cost of quantum ListDecoding search is $2^{(0.284\ldots+o(1))d}$. If memory is constrained to two dimensions, and qRAM costs $\Theta(N^{1/2})$ Clifford+T gates, the quantum asymptotics match the classical RAM asymptotics.

**Quantum sampling routines.** We have assumed that $\mathbf{D}$ in Section 4.1 (the uniform sampling subroutine in Grover's algorithm) is implemented using parallel $\mathbf{H}$ gates. This is the smallest possible circuit that might implement $\mathbf{D}$, and may be a significant underestimate. In Line 12 of Algorithm 4 we must construct a superposition (ideally uniform) over $\{k : v_k \in L_{F,j}\}$. The set $L_{F,j}$ is presented as a disjoint union of smaller sets. Copying the elements of these smaller sets to a flat array would be more expensive than our estimate for the cost of search. While we do not expect the cost of sampling near uniformly from $L_{F,j}$ to be large, it could easily exceed the cost of `popcount`.

### 7.2 Relevance to SVP

The NNS algorithms that we have analysed are closely related to lattice sieves for SVP. While the asymptotic cost of NNS algorithms are often used as a proxy for the asymptotic cost of solving SVP, we caution the reader against making this comparison in a non-asymptotic setting. On the one hand, our estimates might lead one to underestimate the cost of solving SVP:

- the costs given in Figure 2 represent one iteration of NNS within a sieve, while sieve algorithms make $\mathsf{poly}(d)$ iterations;
- the costs given in Figure 2 do not account for all of the subroutines within each NNS algorithm.

On the other hand, our estimates might lead one to overestimate the cost of solving SVP:

- it is a mistake to conflate the cost of NNS in dimension $d$ with the cost of SVP in dimension $d$. The "dimensions for free" technique of [17] can be used to solve SVP in dimension $d$ by calling an NNS routine polynomially many times in dimension $d' < d$. Our analysis seamlessly applies to dimension $d'$;
- there are heuristics that exploit structure present in applications to SVP not captured in our general setting, e.g. the vector space structure allowing both $\pm u$ to be tested for the cost of $u$, and keeping the vectors sorted by length.

### 7.3 Future work

The sieving techniques considered here are not exhaustive. While it would be relatively easy to adapt our software to other 2-sieves, like the cross polytope sieve [10], future work might consider $k$-sieves such as [7, 30].

Future work might also address the barriers to a quantum advantage discussed in Section 7.1. Two additional barriers are worth mentioning here. First, as Grover search does not parallelise well, one might consider depth restrictions for classical and quantum circuits. Second, our estimates might be refined by including some of the classical subroutines, present in both the classical and quantum variants of the same sieve, that we have ignored, e.g. the cost of sampling lattice vectors or the cost of list-decoding in Algorithm 4. Any cost increase will reduce the range of cryptanalytically relevant dimensions, giving fewer dimensions to overcome quantum overheads.

Finally, our estimates should be checked against experiments. Our analysis of Algorithm 3 recommends a database of size $N(d) \approx 2/C_d(\pi/3)$, while the largest sieving experiments to date [2] runs Algorithm 3 with a database of size $N'(d) = 3.2 \cdot 2^{0.2075d}$ up to dimension $d = 127$. There is a factor of 8 gap between $N'(127)$ and $N(127)$. A factor of two can be explained by the fact that [2] treats each database entry $u$ as $\pm u$. It is possible that the remaining factor of four can be explained by the other heuristics used in [2]. As $d$ increases, $N(d)$ and $N'(d)$ continue to diverge, so future work could attempt to determine more accurately the required list size.

## References


1. Ajtai, M., Kumar, R., Sivakumar, D.: A sieve algorithm for the shortest lattice vector problem. In: 33rd ACM STOC. pp. 601–610. ACM Press (Jul 2001)
2. Albrecht, M.R., Ducas, L., Herold, G., Kirshanova, E., Postlethwaite, E.W., Stevens, M.: The general sieve kernel and new records in lattice reduction. In: Ishai, Y., Rijmen, V. (eds.) EUROCRYPT 2019, Part II. LNCS, vol. 11477, pp. 717–746. Springer, Heidelberg (May 2019)
3. Alkim, E., Ducas, L., Pöppelmann, T., Schwabe, P.: Post-quantum key exchange - A new hope. In: Holz, T., Savage, S. (eds.) USENIX Security 2016. pp. 327–343. USENIX Association (Aug 2016)
4. Amy, M., Matteo, O.D., Gheorghiu, V., Mosca, M., Parent, A., Schanck, J.M.: Estimating the cost of generic quantum pre-image attacks on SHA-2 and SHA-3. In: Avanzi, R., Heys, H.M. (eds.) SAC 2016. LNCS, vol. 10532, pp. 317–337. Springer, Heidelberg (Aug 2016)
5. Arunachalam, S., Gheorghiu, V., Jochym-O'Connor, T., Mosca, M., Srinivasan, P.V.: On the robustness of bucket brigade quantum ram. New Journal of Physics 17(12), 123010 (2015), http://stacks.iop.org/1367-2630/17/i=12/a=123010
6. Babbush, R., Gidney, C., Berry, D.W., Wiebe, N., McClean, J., Paler, A., Fowler, A., Neven, H.: Encoding electronic spectra in quantum circuits with linear T complexity. Phys. Rev. X 8, 041015 (Oct 2018), https://link.aps.org/doi/10.1103/PhysRevX.8.041015
7. Bai, S., Laarhoven, T., Stehlé, D.: Tuple lattice sieving. LMS Journal of Computation and Mathematics 19(A), 146–162 (2016)

8. Becker, A., Ducas, L., Gama, N., Laarhoven, T.: New directions in nearest neighbor searching with applications to lattice sieving. In: Krauthgamer, R. (ed.) 27th SODA. pp. 10–24. ACM-SIAM (Jan 2016)
9. Becker, A., Gama, N., Joux, A.: Speeding-up lattice sieving without increasing the memory, using sub-quadratic nearest neighbor search. Cryptology ePrint Archive, Report 2015/522 (2015), http://eprint.iacr.org/2015/522
10. Becker, A., Laarhoven, T.: Efficient (ideal) lattice sieving using cross-polytope LSH. In: Pointcheval, D., Nitaj, A., Rachidi, T. (eds.) AFRICACRYPT 16. LNCS, vol. 9646, pp. 3–23. Springer, Heidelberg (Apr 2016)
11. Bernstein, D.J.: Cost analysis of hash collisions: Will quantum computers make sharcs obsolete? Workshop Record of SHARCS'09: Special-purpose Hardware for Attacking Cryptographic Systems (2009), http://cr.yp.to/papers.html#collisioncost
12. Boyer, M., Brassard, G., Høyer, P., Tapp, A.: Tight bounds on quantum searching. Fortschritte der Physik 46(4-5), 493–505 (1998), https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291521-3978%28199806%2946%3A4/5%3C493%3A%3AAID-PROP493%3E3.0.CO%3B2-P
13. Brassard, G., Hoyer, P., Mosca, M., Tapp, A.: Quantum amplitude amplification and estimation. Contemporary Mathematics 305, 53–74 (2002), https://arxiv.org/abs/quant-ph/0005055
14. Brassard, G., Høyer, P., Tapp, A.: Quantum cryptanalysis of hash and claw-free functions. SIGACT News 28(2), 14–19 (Jun 1997), http://doi.acm.org/10.1145/261342.261346
15. Charikar, M.: Similarity estimation techniques from rounding algorithms. In: 34th ACM STOC. pp. 380–388. ACM Press (May 2002)
16. Cuccaro, S.A., Draper, T.G., Kutin, S.A., Moulton, D.P.: A new quantum ripple-carry addition circuit (2004), arXiv:quant-ph/0410184
17. Ducas, L.: Shortest vector from lattice sieving: A few dimensions for free. In: Nielsen, J.B., Rijmen, V. (eds.) EUROCRYPT 2018, Part I. LNCS, vol. 10820, pp. 125–145. Springer, Heidelberg (Apr / May 2018)
18. Fitzpatrick, R., Bischof, C.H., Buchmann, J., Dagdelen, Ö., Göpfert, F., Mariano, A., Yang, B.Y.: Tuning GaussSieve for speed. In: Aranha, D.F., Menezes, A. (eds.) LATINCRYPT 2014. LNCS, vol. 8895, pp. 288–305. Springer, Heidelberg (Sep 2015)
19. Fowler, A.G., Mariantoni, M., Martinis, J.M., Cleland, A.N.: Surface codes: Towards practical large-scale quantum computation. Phys. Rev. A 86, 032324 (Sep 2012), https://link.aps.org/doi/10.1103/PhysRevA.86.032324
20. Gidney, C., Ekerå, M.: How to factor 2048 bit rsa integers in 8 hours using 20 million noisy qubits (2019), https://arxiv.org/abs/1905.09749, arXiv:1905.09749
21. Giovannetti, V., Lloyd, S., Maccone, L.: Quantum random access memory. Phys. Rev. Lett. 100, 160501 (Apr 2008), http://link.aps.org/doi/10.1103/PhysRevLett.100.160501
22. Gottesman, D.: Fault-tolerant quantum computation with constant overhead (2013), https://arxiv.org/abs/1310.2984, arXiv:1310.2984
23. Grassl, M., Langenberg, B., Roetteler, M., Steinwandt, R.: Applying grover's algorithm to AES: Quantum resource estimates. In: Takagi, T. (ed.) Post-Quantum Cryptography - 7th International Workshop, PQCrypto 2016. pp. 29–43. Springer, Heidelberg (2016)
24. Grover, L., Rudolph, T.: How significant are the known collision and element distinctness quantum algorithms. Quantum Info. Comput. 4, 201–206 (May 2004)

25. Grover, L.K.: Quantum mechanics helps in searching for a needle in a haystack. Phys. Rev. Lett. 79, 325–328 (Jul 1997), http://link.aps.org/doi/10.1103/PhysRevLett.79.325
26. Häner, T., Roetteler, M., Svore, K.M.: Factoring using 2n + 2 qubits with toffoli based modular multiplication. Quantum Info. Comput. 17(7-8), 673–684 (Jun 2017), http://dl.acm.org/citation.cfm?id=3179553.3179560
27. Horsman, C., Fowler, A.G., Devitt, S., Meter, R.V.: Surface code quantum computing by lattice surgery. New Journal of Physics 14(12), 123011 (2012), http://stacks.iop.org/1367-2630/14/i=12/a=123011
28. Jaques, S., Naehrig, M., Roetteler, M., Virdia, F.: Implementing grover oracles for quantum key search on aes and lowmc. Cryptology ePrint Archive, Report 2019/1146 (2019), https://eprint.iacr.org/2019/1146
29. Jaques, S., Schanck, J.M.: Quantum cryptanalysis in the RAM model: Claw-finding attacks on SIKE. In: Boldyreva, A., Micciancio, D. (eds.) CRYPTO 2019, Part I. LNCS, vol. 11692, pp. 32–61. Springer, Heidelberg (Aug 2019)
30. Kirshanova, E., Mårtensson, E., Postlethwaite, E.W., Moulik, S.R.: Quantum algorithms for the approximate k-list problem and their application to lattice sieving. In: Galbraith, S.D., Moriai, S. (eds.) Advances in Cryptology – ASIACRYPT 2019. pp. 521–551. Springer International Publishing, Cham (2019)
31. Klein, P.N.: Finding the closest lattice vector when it's unusually close. In: Shmoys, D.B. (ed.) 11th SODA. pp. 937–941. ACM-SIAM (Jan 2000)
32. Kuperberg, G.: Another subexponential-time quantum algorithm for the Dihedral Hidden Subgroup Problem. In: Theory of Quantum Computation, Communication and Cryptography – TQC 2013. pp. 20–34. LIPIcs 22 (2013), http://drops.dagstuhl.de/opus/volltexte/2013/4321
33. Laarhoven, T.: Sieving for shortest vectors in lattices using angular locality-sensitive hashing. In: Gennaro, R., Robshaw, M.J.B. (eds.) CRYPTO 2015, Part I. LNCS, vol. 9215, pp. 3–22. Springer, Heidelberg (Aug 2015)
34. Laarhoven, T.: Search problems in cryptography: from fingerprinting to lattice sieving. Ph.D. thesis, Department of Mathematics and Computer Science (2 2016), proefschrift
35. Laarhoven, T., Mosca, M., van de Pol, J.: Solving the shortest vector problem in lattices faster using quantum search. In: Gaborit, P. (ed.) Post-Quantum Cryptography. pp. 83–101. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
36. Lindner, R., Peikert, C.: Better key sizes (and attacks) for LWE-based encryption. In: Kiayias, A. (ed.) CT-RSA 2011. LNCS, vol. 6558, pp. 319–339. Springer, Heidelberg (Feb 2011)
37. Maslov, D.: Advantages of using relative-phase toffoli gates with an application to multiple control toffoli optimization. Physical Review A 93(2), 022311 (2016)
38. Matteo, O.D., Gheorghiu, V., Mosca, M.: Fault tolerant resource estimation of quantum random-access memories (2019), arXiv:1902.01329v1
39. Micciancio, D., Regev, O.: Lattice-based cryptography. In: Bernstein, D.J., Buchmann, J., Dahmen, E. (eds.) Post-Quantum Cryptography, pp. 147–191. Springer Berlin Heidelberg, Berlin, Heidelberg (2009), https://doi.org/10.1007/978-3-540-88702-7_5
40. Nguyen, P.Q., Vidick, T.: Sieve algorithms for the shortest vector problem are practical. J. of Mathematical Cryptology 2(2) (2008)
41. Parhami, B.: Efficient hamming weight comparators for binary vectors based on accumulative and up/down parallel counters. IEEE Trans. on Circuits and Systems 56-II(2), 167–171 (2009), https://doi.org/10.1109/TCSII.2008.2010176