Quantum speedup for combinatorial optimization with flat energy landscapes

M. Cain[1], S. Chattopadhyay[1], J.-G. Liu[1,2], R. Samajdar[3,4], H. Pichler[5,6], M. D. Lukin[1]
[1] *Department of Physics, Harvard University, Cambridge, MA 02138, USA*
[2] *Advanced Materials Thrust, Hong Kong University of Science and Technology (Guangzhou), Guangdong 511453, China*
[3] *Department of Physics, Princeton University, Princeton, NJ 08544, USA*
[4] *Princeton Center for Theoretical Science, Princeton University, Princeton, NJ 08544, USA*
[5] *Institute for Theoretical Physics, University of Innsbruck, Innsbruck A-6020, Austria*
[6] *Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Innsbruck A-6020, Austria*
(Dated: July 10, 2023)

Designing quantum algorithms with a speedup over their classical analogs is a central challenge in quantum information science. Motivated by recent experimental observations of a superlinear quantum speedup in solving the Maximum Independent Set problem on certain unit-disk graph instances [Ebadi *et al.*, Science **376**, 6598 (2022)], we develop a theoretical framework to analyze the relative performance of the optimized quantum adiabatic algorithm and a broad class of classical Markov chain Monte Carlo algorithms. We outline conditions for the optimized adiabatic algorithm to achieve a quadratic speedup on hard problem instances featuring flat low-energy landscapes and provide example instances with either a quantum speedup or slowdown. We then introduce an additional local Hamiltonian with no sign problem to the optimized adiabatic algorithm to achieve a quadratic speedup over a wide class of classical simulated annealing, parallel tempering, and quantum Monte Carlo algorithms in solving these hard problem instances. Finally, we use this framework to analyze the experimental observations.

1. INTRODUCTION

Combinatorial optimization problems have wide-ranging applications in science and technology [1]. They are foundational to modern computer science because they encompass NP-hard problems which cannot be solved efficiently by known algorithms. A central challenge in quantum information science is to understand when quantum algorithms can outperform their classical counterparts in solving such NP-hard combinatorial optimization problems [2, 3]. The most general classical combinatorial optimization algorithms seek to minimize a cost function over a set of bit strings. This includes broad classes of Markov chain Monte Carlo algorithms such as simulated annealing (SA) and parallel tempering [4], which simulate cooling to low-temperature states of a classical Hamiltonian encoding the cost function.

Quantum adiabatic algorithms (QAAs) [5] can be viewed as quantum analogs of such general-purpose classical solvers. QAA prepares low-energy states of a classical cost Hamiltonian [6] by adiabatic evolution. The relative performance of QAA and SA is not generically well understood beyond numerical studies [7–9], and theoretical examples of quantum speedup are either restricted to specifically constructed problem instances [10] or require unphysical Hamiltonians [11–13]. However, unlike other quantum algorithms that are known to generically achieve a quadratic speedup over SA [14–16], QAA can be studied experimentally on existing quantum devices. Although early experimental implementations of QAA lacked the many-body coherence believed to be necessary for quantum speedup [17–22], a recent study using a programmable Rydberg atom array [23] observed a superlinear speedup over SA in solving certain hard instances of the NP-hard Maximum Independent Set problem on unit-disk graphs.

Motivated by these experimental results, in this work we develop a theoretical framework to analyze the relative performance of optimized QAA and several classical Markov chain Monte Carlo algorithms. Specifically, we focus on problem instances with flat energy landscapes comprised of many suboptimal configurations of the same cost, over which algorithms must search to find the optimal solution. We show that the QAA's performance is determined by (de)localization of the low-energy eigenstates of the adiabatic Hamiltonian in configuration space: when the low-energy eigenstates are delocalized, and the quantum evolution is optimized to maintain adiabaticity, QAA achieves a quadratic speedup over a wide class of SA and parallel tempering algorithms. To illustrate these concepts, we provide examples of problem instances that feature either a quantum speedup or slowdown depending on the localization of the low-energy eigenstates.

Having developed this framework, we then use it to introduce a modification of QAA that achieves a quadratic speedup over SA and parallel tempering on certain hard Maximum Independent Set problem instances. Importantly, our algorithm only uses local Hamiltonians with no sign problem, meaning that all the off-diagonal matrix elements are non-positive. While QAA Hamiltonians without a sign problem are typically amenable to simulation with quantum Monte Carlo (QMC) – and many prior speedups over SA in this setting have indeed been recovered by QMC [13, 24] – we nevertheless show that our algorithm maintains a quadratic speedup over a wide class of path-integral QMC algorithms. Finally, we apply these techniques to interpret the experimental observations reported in Ref. [23]. We identify instances with better-than-classical performance due to either delocal-

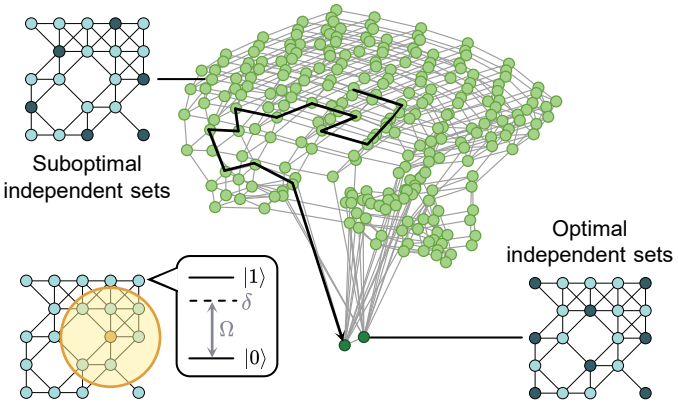

FIG. 1. Flat energy landscapes in combinatorial optimization. The goal of the Maximum Independent Set problem is to find the largest independent sets of a graph (e.g., the dark blue vertices, bottom right) among many suboptimal independent sets (top left). The dynamics of SA on this problem can be visualized by a configuration graph (center), where vertices represent individual independent sets and edges link sets connected by an SA update. SA algorithms randomly walk (black lines) between suboptimal independent sets of the same size (light green vertices) until finding an optimal independent set (dark green vertices). We study QAA's performance on unit-disk graphs (bottom left), where vertices are connected within a unit radius (yellow circle). Each vertex is associated with a qubit with a time-dependent drive $\Omega(t)$ and detuning $\delta(t)$.

ization or favorable localization of the low-energy eigenstates. Instances with worse-than-classical performance can be explained by unfavorable localization of the eigenstates, as introduced by Ref. [25].

Before proceeding, we note that state-of-the-art classical heuristic algorithms specialized to the Maximum Independent Set problem can outperform SA (e.g., [26]). These algorithms accelerate the computation by exploiting the problem-specific graph structure. In contrast, SA is a general-purpose solver that only uses the energy of a configuration in decision-making to prepare the Gibbs distribution of the cost Hamiltonian. Similarly, QAA only takes in the cost Hamiltonian as an input, and prepares its ground state by adiabatic evolution. We will restrict our analysis to the case where the QAA evolution is slow enough to maintain adiabaticity, and the SA evolution is long enough to equilibrate to the Gibbs distribution. Running these algorithms at short, *diabatic* timescales and exploring shortcuts to adiabaticity is of independent interest [27–30].

A. Maximum Independent Set

Throughout this work, we focus on the Maximum Independent Set problem, a paradigmatic NP-hard optimization problem that involves finding the largest independent set of a graph. An independent set is a subset of vertices where no two vertices are connected by an edge. The largest independent set for a graph $G = (V, E)$ with n vertices is a configuration $|z\rangle \in \{|0\rangle, |1\rangle\}^n$ minimizing $H_{\text{cost}}(z) = \langle z| H_{\text{cost}} |z\rangle$ for $\delta > 0$, where

$$H_{\text{cost}} = -\delta \sum_{u\in V} n_u + U \sum_{(u,v)\in E} n_u n_v \tag{1}$$

is the classical cost Hamiltonian. Here, $n_u \equiv |1_u\rangle \langle 1_u|$, and $|1_u\rangle$ ($|0_u\rangle$) denotes that vertex u is present (absent) in the independent set. $U \gg |\delta|$ penalizes edges that violate the independent set constraint. We focus primarily on unit-disk graphs, where edges connect vertices within a unit radius on a two-dimensional plane. These graphs naturally model problems with geometrically local connectivity, such as wireless communication networks [31].

The Maximum Independent Set problem on unit-disk graphs can be naturally encoded in Rydberg atom arrays as follows [32]. Every vertex is associated with an atomic qubit placed on a square grid at position r_u (Fig. 1). The full system is described by the many-body Hamiltonian $H = H_{\text{Ryd}} - H_q$, where

$$H_q = \Omega \sum_{u\in V} |1_u\rangle \langle 0_u| + \text{h.c.}, \tag{2}$$

$$H_{\text{Ryd}} = -\delta \sum_{u\in V} n_u + \sum_{u,v} V_{uv} n_u n_v, \tag{3}$$

and $\Omega(t) > 0$ and $\delta(t)$ are time-dependent energies controlled by a coherent laser drive. The distance-dependent Rydberg blockade interaction energy $V_{uv} \sim 1/|r_u - r_v|^6$ makes simultaneous excitation of two atoms in the Rydberg state $|1_u 1_v\rangle$ within a certain radius energetically unfavorable, mimicking U in Eq. (1). In practice, the blockade radius is chosen to encompass nearest and next-nearest neighbors on the grid.

2. SIMULATED ANNEALING RUNTIME ON HARD INSTANCES

We first characterize fundamentally hard graph instances for SA to find the largest independent set. SA stochastically samples spin configurations from the thermal Gibbs distribution π of H_{cost} at a low temperature $1/\beta$. We consider any Metropolis-Hastings SA algorithm [33, 34] in which the probability $P_{z,z'}$ to update $|z\rangle$ to $|z'\rangle$ satisfies the detailed balance condition,

$$P_{z,z'}\pi_z = P_{z',z}\pi_{z'} \qquad \pi_z = e^{-\beta H_{\text{cost}}(z)}/\mathcal{Z}_\beta, \tag{4}$$

where π_z is the Gibbs population of $|z\rangle$ and $\mathcal{Z}_\beta$ is the partition function. We allow the update rule to be arbitrarily non-local.

Within this general setting, we find that flat energy landscapes, defined as many suboptimal independent sets of the same size with few larger independent sets, form a fundamental obstacle for SA to find the solution [23, 35]. Figure 1 visualizes the flat energy landscape of an example unit-disk graph as a *configuration graph*, where vertices represent independent sets and edges represent SA

updates (here, spin-exchange and spin-flip operations). This instance has many suboptimal independent sets of size $\alpha-1$ and few optimal largest independent sets of size α. The SA dynamics, governed by Eq. (4), are dominated by a random walk among the suboptimal, equal-energy configurations, reminiscent of unstructured search for the optimal solutions. Therefore, we expect the SA runtime to go like the inverse rate $\simeq D_{\alpha-1}/D_\alpha$ of randomly choosing an optimal independent set, where D_b is the number of independent sets of size b.

We now formalize this intuition and describe a lower bound on the SA runtime $\tau_{\mathrm{SA}}(\varepsilon)$. $\tau_{\mathrm{SA}}(\varepsilon)$ is a proxy for the time needed for SA to find an optimal solution. In particular, it given by the SA *mixing time*: the number of proposed updates, normalized by n, needed to prepare the Gibbs distribution with total variation distance $\varepsilon < 1/2$ starting from any initial configuration [36]. As the temperature $1/\beta \to 0$, the Gibbs distribution approaches the uniform mixture of optimal configurations. Thus, if the time for SA to equilibrate amongst the optimal configurations is small compared to the time to find an optimal configuration, we expect $\tau_{\mathrm{SA}}(\varepsilon)$ to represent the time to find a solution. We confirm that this is the case in Appendix E 2, because the optimal configurations are well-connected under spin-exchange updates.

We prove the lower bound on $\tau_{\mathrm{SA}}(\varepsilon)$ in Appendix A 1 by relating $\tau_{\mathrm{SA}}(\varepsilon)$ to the inverse spectral gap $\Delta_{\mathrm{SA}}^{-1}$ of the SA Markov chain transition matrix $P = (P_{z,z'})$ [37]. We then use the Cheeger inequality [38] to relate Δ_{SA} to the flow of population in the Gibbs distribution from independent sets of size $\leq b-1$ to size $\geq b$ during a single SA update. This flow is proportional to D_b/D_{b-1}, which gives us

$$\tau_{\mathrm{SA}}(\varepsilon) \geq \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2nk} \max \frac{D_{b-1}}{D_b}, \tag{5}$$

where k is the maximum number of spins altered during a proposed update [39]. We numerically find in Appendix B that $\max_b(D_{b-1}/D_b)$, and therefore the SA runtime, grows exponentially in $\sqrt{n}$. Moreover, we demonstrate that a similar bound holds for a wide class of parallel tempering algorithms in Appendix A 2. As our proofs are framed in terms of a generic discrete cost function, they also apply to combinatorial optimization problems beyond Maximum Independent Set.

Figure 2 shows the time for an optimized SA algorithm [23] to find an optimal solution with probability 3/4 against Eq. (5), which we compute via a tensor-network algorithm [40]. We plot the data for the top 5% hardest unit-disk graphs maximizing Eq. (5) within each system size ($n = 39-460$, see Appendix B), omitting a small fraction (0.9%) of instances for which the SA runtime is too long to collect sufficient statistics. The strong linear relationship in Fig. 2 confirms that the SA runtime is dominated by unstructured search over flat energy landscapes. Furthermore, it indicates that $\tau_{\mathrm{SA}}(\varepsilon)$ is representative of the time to find an optimal solution.

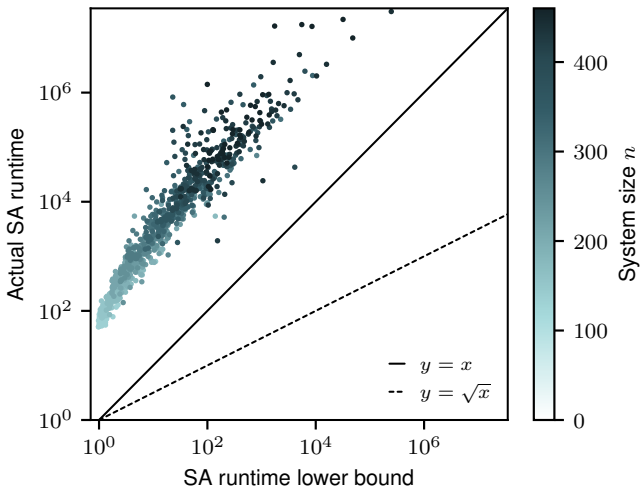

FIG. 2. Flat energy landscapes determine SA runtime. The actual SA runtime to find an optimal solution with probability 3/4 is linearly related to the analytic SA runtime lower bound in Eq. (5), confirming that SA runtime is dominated by overcoming flat energy landscapes.

3. INSTANCE-BY-INSTANCE PERFORMANCE OF QAA

We now establish conditions for which QAA outperforms SA on such hard instances. QAA prepares the ground state of H_{cost} by adiabatic evolution under

$$H_{\mathrm{QAA}} = H_{\mathrm{cost}} - H_q, \tag{6}$$

where the energies $\Omega(t), \delta(t)$ [Eqs. (1) and (2)] vary in time as shown in Fig. 3(a). In particular, we assume that $\Omega(t), \delta(t)$ are optimized to minimize the evolution time while maintaining adiabaticity near the minimum energy gap Δ_{QAA} between the ground and first-excited states of the dominant avoided level crossing, so the runtime of QAA goes as $\Delta_{\mathrm{QAA}}^{-1}$ (specifically, $|dH/dt| \propto \Delta_{\mathrm{QAA}}^2$ at the avoided crossing location $(\Omega/\delta)_\star$, see Refs. [11, 41] and Sec. 6 for further discussion). We will show that Δ_{QAA} is controlled by the properties of two states, $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$, which approximate the ground and first-excited states at this avoided crossing, as shown in Fig. 3(b). We analyze three qualitatively distinct behaviors for $|\mathcal{G}\rangle, |\mathcal{E}\rangle$, which we term *delocalized, favorably localized*, and *unfavorably localized*. The former two result in a speedup over SA, while the latter causes a slowdown.

As argued in Appendix C 3, generically, the avoided level crossing occurs near the end of the ramp, at $(\Omega/\delta)_\star \ll 1$. We will show later that $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ can be computed at leading order in Ω/δ as non-negative superpositions of optimal and suboptimal independent sets,

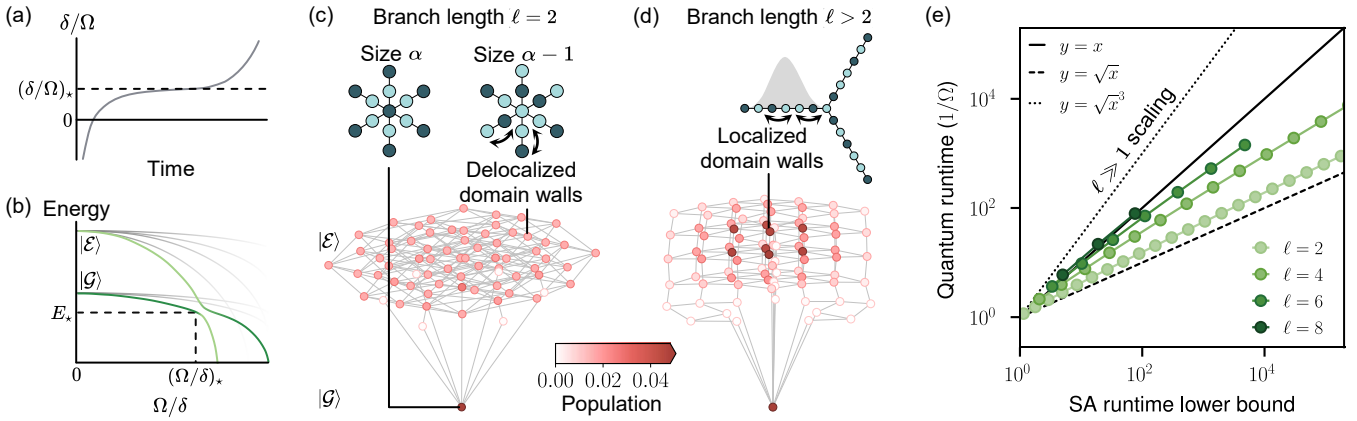

FIG. 3. Eigenstate localization determines QAA runtime. (a) The optimized QAA runtime is proportional to $\Delta_{\mathrm{QAA}}^{-1}$ when the system Hamiltonian changes slowly at $(\delta/\Omega)_\star$, the location of the avoided level crossing. (b) Δ_{QAA} can be computed perturbatively when $(\Omega/\delta)_\star \ll 1$ from Eq. (8), which describes the coupling under the Hamiltonian between the estimated eigenstates $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ involved in the avoided level crossing. (c) The star graph with n_b branches of even length ℓ has a unique optimal independent set of size α with the central vertex in the independent set (top left). It has approximately $(\ell/2+1)^{n_b}$ suboptimal independent sets of size $\alpha - 1$ with the central vertex absent (top right), corresponding to all possible locations of a domain wall on each branch. When $\ell = 2$, the two possible domain wall locations on each branch are equally energetically favored, causing $|\mathcal{E}\rangle$ to delocalize over all domain wall locations. (d) When $\ell > 2$, $|\mathcal{E}\rangle$ localizes around configurations with the domain walls near the center of each branch. (e) QAA has a quadratic speedup in runtime over SA as a function of n_b for the delocalized case of $\ell = 2$. As ℓ increases, $|\mathcal{E}\rangle$ localizes away from $|\mathcal{G}\rangle$, causing SA to outperform QAA when $\ell \gg 1$.

respectively:

$$|\mathcal{G}\rangle = \sum_{z:H_{\mathrm{cost}}(z)=-\delta\alpha} \sqrt{\mathcal{G}_z}\,|z\rangle\,,$$
$$|\mathcal{E}\rangle = \sum_{z:H_{\mathrm{cost}}(z)=-\delta(\alpha-1)} \sqrt{\mathcal{E}_z}\,|z\rangle\,. \tag{7}$$

In the examples we consider, $|\mathcal{E}\rangle$ is a superposition of independent sets of size $\alpha - 1$, though our arguments can be generalized when $|\mathcal{E}\rangle$ is a superposition of smaller independent sets (see Appendix C 3). We can estimate Δ_{QAA} in powers of $(\Omega/\delta)_\star$ as the coupling between $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ [25, 42–45],

$$\tilde{\Delta}_{\mathrm{QAA}} = 2\left|\sum_{l=0}^{\infty} \langle\mathcal{E}|\left(H_q \frac{Q}{E_\star - H_{\mathrm{cost}}}\right)^l H_q\,|\mathcal{G}\rangle\right|, \tag{8}$$

where $Q = \mathbb{1} - |\mathcal{E}\rangle\langle\mathcal{E}| - |\mathcal{G}\rangle\langle\mathcal{G}|$. In Appendix C 1, we derive a bounded proportionality factor relating Δ_{QAA} and $\tilde{\Delta}_{\mathrm{QAA}}$. We note that these results provide a perturbative approach to *exactly* compute Δ_{QAA} and are thus of broader utility and interest beyond the specifics of the problem considered here.

Per Eq. (8), $\tilde{\Delta}_{\mathrm{QAA}}$ is determined by the distribution of wavefunction amplitudes in $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$. At each order l in $(\Omega/\delta)_\star$, factors of H_q generate $l+1$ spin flips to connect pairs of configurations in $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$. The leading-order coupling between two configurations $|z\rangle$ and $|z'\rangle$ within Hamming distance $l+1$ goes like $\sqrt{\mathcal{G}_z\mathcal{E}_{z'}}(\Omega/\delta)_\star^l$. This coupling is enhanced for sets with larger amplitude but is suppressed exponentially in l. This intuition leads us to distinguish between problem instances where $\mathcal{G}_z$ and $\mathcal{E}_z$ are localized on comparatively few sets, and those where they are distributed more evenly among all sets. We refer to instances where $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ localize on sets sufficiently far apart in Hamming distance such that QAA suffers a slowdown relative to SA ($\tilde{\Delta}_{\mathrm{QAA}} \ll D_\alpha/D_{\alpha-1}$) as *unfavorably localized* [25]. By contrast, on *favorably localized* instances, $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ localize at small Hamming distances, such that QAA has speedup over SA ($\tilde{\Delta}_{\mathrm{QAA}} \gg D_\alpha/D_{\alpha-1}$). Several previous notable instances where QAA has an exponential speedup [10] or slowdown [30] fall into these two categories.

QAA also outperforms SA on *delocalized* instances, where the amplitudes $\sqrt{\mathcal{G}_z}$ and $\sqrt{\mathcal{E}_z}$ are close to uniform. Suppose that $|\mathcal{G}\rangle = |S_\alpha\rangle$ and $|\mathcal{E}\rangle = |S_{\alpha-1}\rangle$, where

$$|S_b\rangle = \frac{1}{\sqrt{D_b}} \sum_{z:H_{\mathrm{cost}}(z)=-\delta b} |z\rangle\,. \tag{9}$$

The lowest-order ($l = 0$) contribution to Eq. (8) is then

$$\tilde{\Delta}_{\mathrm{QAA}} = 2|\langle S_{\alpha-1}|\,H_q\,|S_\alpha\rangle\,| = \frac{2}{\sqrt{D_{\alpha-1}D_\alpha}} \sum_{z:H_{\mathrm{cost}}(z)=-\delta\alpha} \Omega\alpha$$
$$= 2\Omega\alpha\sqrt{\frac{D_\alpha}{D_{\alpha-1}}}. \tag{10}$$

Due to coherent enhancement in the coupling, here, the QAA runtime $\tilde{\Delta}_{\mathrm{QAA}}^{-1}$ is quadratically smaller than the SA runtime [Eq. (5)] up to polynomial factors in n. This is reminiscent of the adiabatic version of Grover's search [11], which has a similar quadratic speedup over randomly guessing in $\{|0\rangle, |1\rangle\}^n$ for optimal solutions.

However, we emphasize that the runtimes of QAA and SA in Eqs. (10) and (5), respectively, are asymptotically faster than Grover's search, because they search only among near-optimal configurations for the largest independent set.

A. Determining eigenstate localization

Given a problem instance, we can determine $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ by performing second-order perturbation theory in the degenerate manifolds of H_{cost}. For simplicity, we take the energy penalty on independent set violations $U \to \infty$, so that each degenerate manifold contains independent sets of the same size. The perturbed eigenstates (energy shifts) are the eigenvectors (eigenvalues) of the matrix

$$H^{(2)} = -\frac{\Omega^2}{\delta}\left(H_{se} + \sum_{u\in V}\left[n_u - (\mathbb{1}-n_u)\prod_{(u,v)\in E}(\mathbb{1}-n_v)\right]\right), \tag{11}$$

where H_{se} is the spin-exchange Hamiltonian,

$$H_{se} = \sum_{(u,v)\in E} \sigma_u^+\sigma_v^- + \sigma_u^-\sigma_v^+, \tag{12}$$

$\sigma_u^+ = |1_u\rangle\langle 0_u|$, and $\sigma_u^- = |0_u\rangle\langle 1_u|$. $|\mathcal{G}\rangle$ is the ground state of $H^{(2)}$ in the $H_{\text{cost}} = -\delta\alpha$ manifold, and $|\mathcal{E}\rangle$ is the ground state of the excited manifold whose energy first intersects $|\mathcal{G}\rangle$ at a finite $(\Omega/\delta)_\star$. As $H^{(2)}$ has no sign problem, $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ have non-negative amplitudes.

We find that first term in Eq. (11), $-(\Omega^2/\delta)H_{se}$, primarily determines the (de)localization of $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$. This is because the second term is uniform within a manifold, and the third term (which counts the number of vertices that can be added to the independent set) is small for near-optimal independent sets. In particular, the expectation value of the third term is at most $-(\Omega^2/\delta)(\alpha - b)$ for an independent set of size b, and is zero when no vertices can be added to a set without removing existing vertices. In order to minimize $-(\Omega^2/\delta)H_{se}$, $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ will thus have larger overlap with independent sets that have more neighboring independent sets connected by spin exchanges in the configuration graph. In contrast, if all configurations in the $H_{\text{cost}} = -\delta b$ manifold have the same degree (number of neighbors), the ground state in that manifold is the delocalized superposition $|S_b\rangle$. This follows from viewing H_{se} as the adjacency matrix of the configuration graph within that manifold, and noting that the principal eigenvector of the adjacency matrix of a graph with regular degree is uniform [46].

B. Delocalization–localization crossover for a family of star graphs

To concretely illustrate these concepts, we explore a family of star graphs, where $|\mathcal{E}\rangle$ can be tuned from delocalized to unfavorably localized. A star graph contains n_b branches of even length ℓ connected by a central vertex. We will compare the QAA and SA runtimes at fixed ℓ as n_b grows. The unique largest independent set includes the central vertex plus alternating vertices on each branch (Fig. 3(c), top left). All but a vanishing fraction of the suboptimal independent sets of size $\alpha - 1$ have the central vertex absent and alternating antiferromagnetic order on the branches, each of which has a single domain wall located in one of $\ell/2+1$ possible positions (Fig. 3(c), top right). The SA runtime is thus exponential in n_b,

$$\tau_{\text{SA}}(\varepsilon) \geq \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2nk}\frac{D_{\alpha-1}}{D_\alpha} \geq \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2nk}(\ell/2+1)^{n_b}. \tag{13}$$

To compute the QAA runtime from Eq. (8), we first calculate $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$. $|\mathcal{G}\rangle$ is the unique largest independent set, and $|\mathcal{E}\rangle$ is the ground state of $H^{(2)}$ in the $H_{\text{cost}} = -\delta(\alpha-1)$ manifold. By the reasoning above, on each branch, $|\mathcal{E}\rangle$ is well-approximated by the ground state of $-(\Omega^2/\delta)H_{se}$, which acts as a one-dimensional hopping Hamiltonian, with open boundary conditions, for each domain wall. Therefore, $|\mathcal{E}\rangle$ is given by

$$\langle x_1 x_2 \ldots x_{n_b}|\mathcal{E}\rangle \simeq \prod_{i=1}^{n_b}\frac{1}{\sqrt{\ell/4+1}}\sin\left(\frac{\pi x_i}{\ell/2+2}\right), \tag{14}$$

where $|x_i\rangle, x_i \in \{1,2,\ldots,\ell/2+1\}$ is the state with the domain wall on the ith branch located between sites $2x_i-2$ and $2x_i-1$ (see Fig. 3(d), top and Appendix D 1). $\tilde{\Delta}_{\text{QAA}}$ can be computed to leading order in Ω/δ from Eq. (8) by connecting $|\mathcal{G}\rangle$ to the set in $|\mathcal{E}\rangle$ with all domain walls adjacent to the central vertex ($x_i = 1$) by flipping the central vertex,

$$\begin{aligned}\tilde{\Delta}_{\text{QAA}} &\simeq 2\Omega|\langle\mathcal{G}|H_q|\mathcal{E}\rangle| \\ &\simeq 2\Omega\left(\frac{1}{\sqrt{\ell/4+1}}\sin\left(\frac{\pi}{\ell/2+2}\right)\right)^{n_b}.\end{aligned} \tag{15}$$

Terms that are higher-order in Ω/δ do not affect the scaling of $\tilde{\Delta}_{\text{QAA}}$ with n_b, as shown in Appendix D 2.

Figure 3(e) plots the numerically computed QAA runtime Δ_{QAA}^{-1} versus the SA runtime lower bound for $\tau_{\text{SA}}(1/4)$ for branch lengths $\ell = 2, 4, 6$, and 8. When $\ell = 2$, $|\mathcal{E}\rangle$ delocalizes evenly among all domain wall configurations (Eq. (14) and Fig. 3(c), bottom), yielding a quadratic quantum speedup because $\Delta_{\text{QAA}}^{-1} = 2\Omega\sqrt{2}^{n_b} \lesssim \sqrt{\tau_{\text{SA}}(\varepsilon)}$. As ℓ increases, according to Eq. (14), $|\mathcal{E}\rangle$ unfavorably localizes away from $|\mathcal{G}\rangle$, on sets with the domain wall located near the center of each branch (Fig. 3(d), bottom). Expanding Eq. (14) for small angles, we find that this results in a *slowdown* for QAA when $\ell \gg 1$, as $\Delta_{\text{QAA}}^{-1} \simeq \tau_{\text{SA}}(\varepsilon)^{3/2}$.

4. QUANTUM SPEEDUP FROM DELOCALIZATION

A. Quantum speedup over simulated annealing

So far, our results show that the optimized QAA

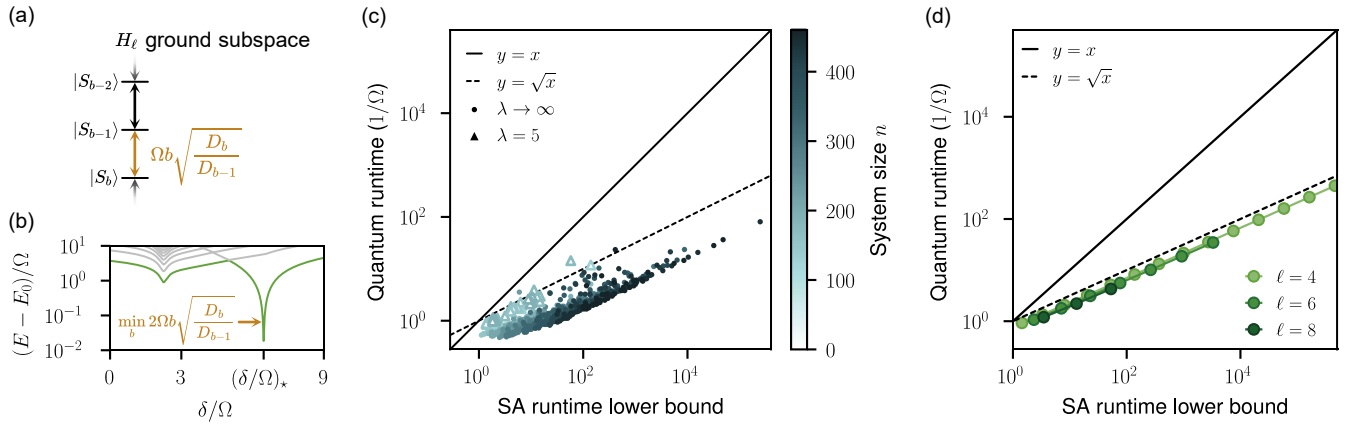

FIG. 4. Quantum speedup over simulated annealing. (a) When $\lambda \to \infty$, the dynamics of the modified QAA [Eq. (16)] are restricted to the degenerate ground states of H_ℓ, which are the uniform superpositions $|S_b\rangle$ of each independent set size b [Eq. (9)]. The matrix elements of H_q (gold) between $|S_b\rangle$ and $|S_{b-1}\rangle$ are coherently enhanced over the analogous rate at which SA transitions from independent sets of size $b-1$ to b. (b) The energy spectrum minus the ground state energy E_0 of an example 720-vertex instance, restricted to the ground subspace of H_ℓ. The minimum gap Δ_{QAA} of the modified QAA is set by the smallest coupling (gold). (c) The modified QAA runtime $\Delta_{\mathrm{QAA}}^{-1}$ scales as the square root of the SA runtime for the same instances as in Fig. 2 when dynamics are restricted to the ground subspace of H_ℓ (circles). The speedup is also obtained for finite $\lambda = 5$ (triangles). (d) The modified QAA obtains a quadratic speedup over SA for the star graphs with branch length $\ell = 4, 6, 8$ when $\lambda = 2.2, 4.1, 6.5$, respectively.

achieves a quadratic speedup over SA when its low-energy eigenstates are delocalized, due to the coherent enhancement of the couplings $\langle S_{b-1}| H_q |S_b\rangle$ in Eq. (10). It is thus natural to ask whether instances with unfavorable localization can be remedied by modifying QAA to force the eigenstates to delocalize. We achieve this result by designing a Hamiltonian H_ℓ whose degenerate ground subspace is spanned by the uniform superpositions $\{|S_b\rangle\}$ $(b = 0, 1, \ldots, \alpha)$, and adding it to the QAA Hamiltonian with a time-independent energy scale λ,

$$H = H_{\mathrm{QAA}} + \lambda H_\ell. \tag{16}$$

In contrast to prior approximate approaches to favoring delocalization [47, 48], this approach provably enforces delocalization under certain conditions on the flat energy landscape, which we will state.

To design H_ℓ, we draw inspiration from the single-particle quantum kinetic energy operator, the ground state of which is maximally delocalized. Since the single-particle kinetic energy is the negative of the continuum Laplacian $-\nabla^2$, we let H_ℓ be the discrete Laplacian of the configuration graph in Fig. 1, restricted to each degenerate manifold of H_{cost}, where vertices represent independent sets and edges represent spin exchanges. The discrete Laplacian is the negative of the adjacency matrix (H_{se}), plus a diagonal term that counts the degree for that configuration, i.e., the number of possible spin exchanges,

$$H_\ell = -H_{se} + \sum_{u\in V} \sum_{(u,v)\in E} n_u(\mathbb{1} - n_v) \prod_{\substack{(y,v)\in E \\ y\neq u}} (\mathbb{1} - n_y), \tag{17}$$

where $G = (V, E)$ is the original problem graph. Crucially, the diagonal term prevents the ground states of H_ℓ from localizing on independent sets with larger degrees on the configuration graph. This differs from the perturbative spin-exchange term in the unmodified QAA Hamiltonian $H^{(2)}$ [Eq. (11)], which energetically favors configurations with more possible spin exchanges. We emphasize that H_ℓ can be efficiently constructed using only local information about the problem graph. For unit-disk graphs embedded on a square grid, the terms in H_ℓ only involve a constant number of spins, which allows for its implementation in near-term experiments.

To develop some intuition, let us first analyze the modified QAA when the energy scale of H_ℓ, λ, is large. If there exists a path between any two configurations in a degenerate manifold under spin exchanges, then each block H_b of $H_\ell = H_0 \oplus H_1 \oplus \ldots \oplus H_\alpha$ has a unique ground state equal to $|S_b\rangle$ with eigenvalue zero [46]. Since the QAA dynamics are restricted to this ground subspace when $\lambda, U \to \infty$, the modified QAA Hamiltonian in Eq. (16) reduces to a one-dimensional tight-binding Hamiltonian,

$$H_{tb} = -\sum_{b=1}^{\alpha} \delta b\, |S_b\rangle \langle S_b| + \Omega b\sqrt{\frac{D_b}{D_{b-1}}}(|S_b\rangle \langle S_{b-1}| + \mathrm{h.c.}), \tag{18}$$

which has an electric field gradient of strength δ and site-dependent couplings $\Omega b\sqrt{D_b/D_{b-1}}$ [see Fig. 4(a)]. If the minimum energy gap Δ_{QAA} of H_{tb} is set by the smallest coupling, as shown in Fig. 4(b) for an example 720-vertex unit-disk graph, then $\Delta_{\mathrm{QAA}}^{-1}$ is quadratically smaller than the SA runtime lower bound. We confirm the trend $\Delta_{\mathrm{QAA}} \simeq \min_b(\Omega b\sqrt{D_b/D_{b-1}})$ numerically for

hundreds of hard instances of the Maximum Independent Set problem on unit-disk graphs in Fig. 4(c). To explain these observations, we show in Appendix E 1 that $\Delta_{\mathrm{QAA}} \simeq \Omega\alpha\sqrt{D_\alpha/D_{\alpha-1}}$ on the vast majority of studied instances, for which the smallest coupling is between independent sets of size $\alpha-1$ and α and the remaining couplings are a smooth function of b. We additionally argue in Appendix E 2 that the same result holds when a small number of configurations within a degenerate manifold are disconnected under spin exchanges, which occurs for a small fraction of instances.

To achieve the quantum speedup in practice, however, $\Delta_{\mathrm{QAA}}^{-1}$ must scale more favorably than the SA runtime when the energy scales of the modified QAA Hamiltonian are measured in units of λ, when λ is the largest energy scale of H. To investigate the scale of λ/Ω required to obtain the quadratic enhancement of Δ_{QAA}, in Fig 4(c) we plot $\Delta_{\mathrm{QAA}}^{-1}$ for the top 1% hardest instances with up to $n=80$ vertices, computed using the density matrix renormalization group method (DMRG) [49, 50]. With the modest overhead of $\lambda/\Omega=5$, we observe a clear quadratic scaling advantage over the SA runtime lower bound in Eq. (5). Furthermore, the modified QAA with $\lambda/\Omega=1$ substantially outperforms the unmodified QAA on the same instances (see Fig. 13(a) of Appendix E 2).

We complement our numerical observations with sufficient, though not necessary, conditions on the λ which yield a quadratic quantum speedup. In Appendix E 2, we show analytically that a sufficient condition for achieving the quadratic enhancement of Δ_{QAA} is $\lambda/\Omega, \lambda/\delta \gtrsim \Delta_{\ell,b}^{-1}, \Delta_{\ell,b-1}^{-1}$, where $\Delta_{\ell,b}, \Delta_{\ell,b-1}$ are the spectral gaps of the delocalizing Hamiltonian H_ℓ restricted to the manifolds b and $b-1$ that share the smallest tight-binding coupling $\min_b(\Omega b\sqrt{D_b/D_{b-1}})$. In Fig. 4(d), we confirm that the modified QAA with $\lambda=\Delta_{\ell,\alpha-1}^{-1}=\mathcal{O}(1)$ has a quadratic speedup for the family of star graphs. We show in Fig. 13(b) of Appendix E 2 that typically $\Delta_{\ell,b}, \Delta_{\ell,b-1} > 1/n$ for the unit-disk graphs we study; accordingly, $\lambda\Delta_{\mathrm{QAA}}^{-1} \sim n\min_b(\Omega b\sqrt{D_{b-1}/D_b})$. Therefore, when $\Delta_{\ell,b}^{-1}, \Delta_{\ell,b-1}^{-1}$ grow at most polynomially in n, the modified QAA's runtime is (sub)exponentially faster than the runtime of Grover's search ($\sqrt{2}^n$) for the hard unit-disk graphs we study: numerically, the SA runtime goes like $c^{\sqrt{n}}$ for some $c\in(1,2)$, whereas the modified QAA runtime is $\sqrt{c}^{\sqrt{n}}$ up to polynomial factors in n (see Appendix B).

B. Quantum speedup over Quantum Monte Carlo

As the modified QAA does not suffer from a sign problem, path-integral QMC can be used to sample independent sets from its thermal Gibbs distribution $\pi_z = \langle z|\, e^{-\beta H}\,|z\rangle\,/\mathcal{Z}_\beta$. In general, path-integral QMC works by stochastically sampling trajectories from a discretized imaginary-time path integral of the partition function $\mathcal{Z}_\beta=\mathrm{Tr}(e^{-\beta H})$. Several prior exponential speedups for QAA over SA have been recovered by sampling from the QMC path integral at low temperatures as the Hamiltonian is varied adiabatically in real time [13, 51]. It is thus natural to ask whether this procedure, also called *simulated quantum annealing*, can match the modified QAA runtime.

In Appendix A 3, we derive a lower bound for the QMC runtime $\tau_{\mathrm{QMC}}(\varepsilon)$ of both the modified and unmodified QAA. Analogous to the SA runtime $\tau_{\mathrm{SA}}(\varepsilon)$, $\tau_{\mathrm{QMC}}(\varepsilon)$ is the number of QMC updates, normalized by n/M, where M is the number of imaginary time slices, needed to sample from π with total variation distance $\varepsilon<1/2$ [52]. We consider any QMC algorithm which alters up to k spins in each imaginary time slice per update, where k is restricted to be constant in n.

Crucial to our argument is the fact that before QMC encounters an independent set $|z\rangle$ with $H_{\mathrm{cost}}(z)\le-\delta b$, it effectively samples from a restricted Hilbert space of only independent sets with $H_{\mathrm{cost}}(z)\ge-\delta(b-1)$. At any point during the adiabatic ramp, we let $H^{(r,b)}$ denote the Hamiltonian in this restricted Hilbert space, with corresponding Gibbs populations $\pi_z^{(r,b)}$. We let $|z_{\max}\rangle$ denote the configuration in this restricted Hilbert space within k spin flips of an independent set of size b with the maximum Gibbs population $\pi_{z_{\max}}^{(r,b)}$. Further, we let $e_{\max}^{(r,b)}=\pi_{z_{\max}}^{(r,b)}D_{b-1}$ describe relative enhancement or suppression of its population compared to the uniform superposition state $|S_{b-1}\rangle$.

Analogous to SA, we then apply the Cheeger inequality to derive an upper bound on the QMC Markov chain spectral gap Δ_{QMC}, which gives a lower bound on $\tau_{\mathrm{QMC}}(\varepsilon)$. This allows us to relate Δ_{QMC} to the flow from populations in the Gibbs distribution of $\pi_z^{(r,b)}$ to independent sets of size $\ge b$. This flow is proportional to $e_{\max}^{(r,b)}D_b/D_{b-1}$, which gives us

$$\tau_{\mathrm{QMC}}(\varepsilon)\ge\frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2nkn^k}\max\frac{D_{b-1}}{e_{\max}^{(r,b)}D_b}. \tag{19}$$

Eq. (19) shows that when the Gibbs distribution of the restricted Hilbert space is delocalized, i.e., when the H_ℓ energy scale λ is sufficiently large, the modified QAA has a quadratic speedup over QMC. In this case, $e_{\max}^{(r,b)}\le1$, so the QMC runtime goes like $\max_b(D_{b-1}/D_b)$, whereas the modified QAA runtime goes like $\max_b\sqrt{D_{b-1}/D_b}$. To match the modified QAA runtime, the restricted Gibbs distribution must be *exponentially* favorably localized, so that $e_{\max}^{(r,b)}=\sqrt{D_{b-1}/D_b}$. In this case, however, we expect the QAA runtime to be similarly enhanced under Eq. (8) due to favorable localization. Thus, QMC does not recover the quadratic speedup due to delocalization, which crucially stems from the quantum coherent enhancement of the coupling $\langle S_{b-1}|\,H_q\,|S_b\rangle$.

5. UNDERSTANDING THE EXPERIMENTAL OBSERVATIONS

We now apply our framework to interpret recent exper-

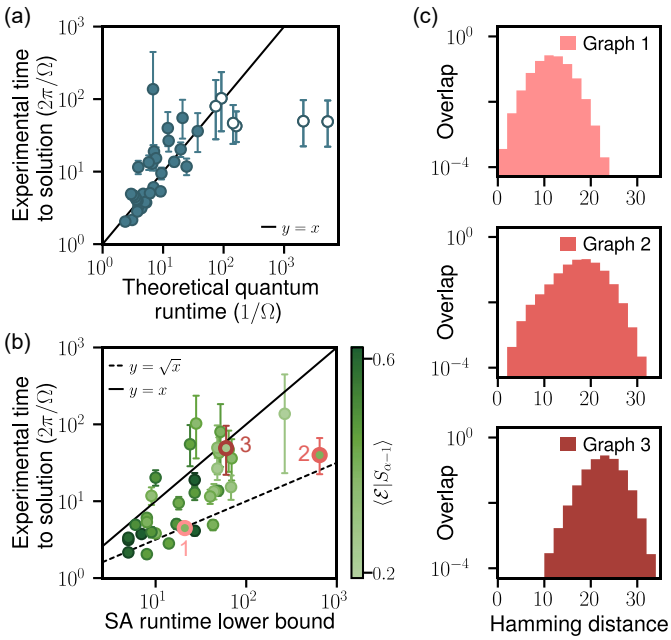

FIG. 5. Analysis of the experimental performance. (a) The experimental optimized time to solution correlates with the theoretical QAA runtime $\Delta_{\mathrm{QAA}}^{-1}$ on instances where the maximum experimental evolution time $T_{\max}$ can resolve the minimum gap ($T_{\max} \leq \Delta_{\mathrm{QAA}}^{-1}$, teal-filled points). Instances for which the evolution time is too short to maintain adiabaticity deviate from the trend ($T_{\max} \geq \Delta_{\mathrm{QAA}}^{-1}$, white points). (b) The experimental time to solution correlates less strongly with the SA runtime lower bound on instances where $|\mathcal{E}\rangle$ is localized (light green points). On the most delocalized instances (dark green points), the QAA runtime is similar to the square root of the SA runtime. (c) We plot the distribution of Hamming distances between $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ for three localized graphs. The pairwise Hamming distances are larger for the instance where QAA performs poorly relative to SA (bottom), and smaller for the instances where QAA outperforms SA (top, middle).

iments on Rydberg atom arrays [23] using the aforementioned hardware-efficient encoding of the Maximum Independent Set problem on unit-disk graphs. Ebadi *et al.* [23] observed that the experimental optimized QAA outperformed SA on certain hard unit-disk graph instances with a large ratio of $D_{\alpha-1}/D_{\alpha}$ ($n = 39-80$). We compute the experimental optimized time to solution as [3]

$$\mathrm{TTS}_{\mathrm{opt}} = \min_{T} \frac{T}{\ln[1-p(T)]}, \tag{20}$$

where $p(T)$ is the probability of QAA finding the optimal solution at evolution time T. In Fig. 5(a), we confirm that $\mathrm{TTS}_{\mathrm{opt}}$ goes like the theoretical runtime $\Delta_{\mathrm{QAA}}^{-1}$ computed numerically for the Rydberg Hamiltonian [Eqs. (2) and (3)].

However, in Fig. 5(b), we find that $\mathrm{TTS}_{\mathrm{opt}}$ correlates less strongly with the SA runtime lower bound. To understand Δ_{QAA}, and therefore the experimental time to solution, we obtain perturbative estimates for the eigenstates at the avoided level crossing, $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$, in the manifold of independent sets of size α and $\alpha-1$, respectively. On the more delocalized instances, $\Delta_{\mathrm{QAA}}^{-1}$ is similar to the square root of the SA runtime (Fig 5(b), dark green points), as expected from perturbation theory [Eq. (10)].

In contrast, for more localized instances (light green points), we find that $\mathrm{TTS}_{\mathrm{opt}}$ is less correlated with the SA runtime. By Eq. (8), we expect Δ_{QAA} to be small when the Hamming distance between $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ is large and $(\Omega/\delta)_\star$ is small, which we verify numerically in Appendix C 4. For illustration, in Fig. 5(c) we examine three localized instances with vastly different SA and QAA runtimes. We plot the distribution of the product of populations $\mathcal{G}_z\mathcal{E}_{z'}$ of spin configurations $|z\rangle, |z'\rangle$ in $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ over their Hamming distances. The instance where SA outperforms QAA is highly localized (bottom, $n = 80$), with large Hamming distances compared to the two other instances where QAA outperforms SA (top and middle, $n = 65$). Due to favorable localization, these instances obtain a significant speedup over SA. Thus, the instance-dependent characteristics of $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ can be used to predict the experimental performance.

6. OUTLOOK

In this work, we have shown that the optimized QAA has a quadratic speedup over a wide class of classical Markov chain algorithms when the low-energy eigenstates are delocalized across a flat energy landscape. To promote delocalization on generic problem instances [25], we modified QAA by adding a local Hamiltonian H_ℓ with no sign problem, with a time-independent energy scale λ. To observe the corresponding quadratic speedup on near-term devices, the algorithm must be efficiently encoded in hardware [53]. The modified QAA is amenable to direct experimental implementation via hybrid digital-analog Trotterized evolution [54], by generating spin-exchange interactions with excitation into S and P Rydberg states or microwave driving [55, 56], and decomposing the diagonal component of H_ℓ into multiqubit controlled phase gates. Local detunings can generate the diagonal component of H_ℓ on certain instances with structured configuration graphs, such as when the suboptimal configurations correspond to the motion of a domain wall [57].

Similar to other problems involving Grover-type quadratic speedups [11, 58], our approach requires optimizing the QAA evolution to maintain adiabaticity. Optimizing QAA evolution in general is an open problem; however, recent work has shown that it is possible to optimize a wide class of QAA algorithms which use the reflection about the uniform superposition state, $\mathbb{1} - \frac{1}{2^n}\sum_{z,z'} |z\rangle\langle z'|$, to drive the evolution instead of H_q [41]. In Appendix E 1 b, we describe approaches to optimizing the modified QAA when $\lambda \to \infty$, which retain a quadratic speedup. Future work could attempt to generalize these results to finite λ. At the same time, one could circumvent the need for optimization by identifying instances with an exponential, rather than quadratic,

speedup over SA. One approach could be to characterize instances where the low-energy eigenstates are favorably localized at small Hamming distance [10]. However, QAA may not generically provide a speedup over QMC on these instances [10, 13, 51]. It remains an open question whether instances exist with an exponential speedup over both SA and QMC, despite optimistic results in the black-box setting [59, 60].

It would also be interesting to extend our results beyond flat energy landscapes to problems with the Overlap Gap Property, whose optimal solutions are provably hard to approximate for large classes of both quantum and classical algorithms [61, 62]. In these instances, independent sets of the same size form "clusters" separated by large Hamming distances. As the clusters are disconnected under spin-exchange operations, they independently delocalize, such that the effective Hamiltonian is a tree-like version of the one-dimensional tight-binding Hamiltonian H_{tb} when $\lambda \to \infty$ [Eq. (18)]. Future work could investigate the modified QAA on instances with the Overlap Gap Property using this framework. Particularly interesting is the prospect of studying QAA performance in the *diabatic* regime, which can outperform both SA and QMC in finding approximate solutions on certain problem instances [29]. Utilizing non-adiabatic phenomena via quantum quench algorithms may provide an alternative mechanism for quantum speedup [28, 30, 63, 64].

7. ACKNOWLEDGEMENTS

We would like to thank Tameem Albash, Lisa Bombieri, Dolev Bluvstein, Sepehr Ebadi, Nicholas Ezzell, Aram Harrow, Marcin Kalinowski, Andrew King, Daniel Lidar, Subir Sachdev, Benjamin Schiffer, Juspreet Singh Sandhu, Lei Wang, and Zhongda Zeng for helpful discussions. This work was supported by the US Department of Energy [DE-SC0021013 and DOE Quantum Systems Accelerator Center (contract no. 7568717)], the Defense Advanced Research Projects Agency (grant no. W911NF2010021), the Army Research Office (grant No. W911NF-21-1-0367), the National Science Foundation, the Harvard-MIT Center for Ultracold Atoms, and the European Research Council (grant No. 101041435). M.C. acknowledges support from Department of Energy Computational Science Graduate Fellowship under Award Number (DESC0020347). S.C. is grateful for support from the NSF under Grant No. DGE-1845298. R.S. is supported by the Princeton Quantum Initiative Fellowship.

Appendix A: Runtime lower bounds for classical Markov chain algorithms

1. Simulated annealing

In this section, we establish a runtime lower bound on all simulated annealing (SA) algorithms using the Metropolis-Hastings update rule. Although we focus on the Maximum Independent Set problem in our proof, we will show that our bound applies to generic combinatorial optimization problems. The goal of SA is to sample from an equilibrium probability distribution π, which we take to be the thermal Gibbs distribution of H_{cost} at temperature $1/\beta$,

$$\pi_z = \frac{e^{-\beta H_{\text{cost}}(z)}}{\mathcal{Z}_\beta}, \qquad \mathcal{Z}_\beta = \sum_b D_b e^{\beta \delta b}, \tag{A1}$$

where π_z is the probability of spin configuration $|z\rangle \in \{|0\rangle, |1\rangle\}^n$, $\mathcal{Z}_\beta$ is the partition function, and D_b is the number of independent sets of size b. SA stochastically updates a spin configuration $|z\rangle$ to $|z'\rangle$ according to the Markov chain transition probabilities $P_{z,z'}$. We consider $P_{z,z'}$ given by the Metropolis-Hastings update rule [33, 34],

$$P_{z,z'} = p_{z,z'} \min\left(1, e^{-\beta[H_{\text{cost}}(z') - H_{\text{cost}}(z)]}\right), \tag{A2}$$

where $p_{z,z'} = p_{z',z}$ is the probability of proposing to update from $|z\rangle$ to $|z'\rangle$, and the remaining factor is the probability of accepting the proposed update. One can check that for $p_{z,z'} = p_{z',z}$, the update rule satisfies the detailed balance condition,

$$P_{z,z'}\pi_z = P_{z',z}\pi_{z'}. \tag{A3}$$

The *mixing time* of SA is defined as the minimum number of proposed updates per spin to prepare the Gibbs distribution with error (measured in total variation distance, see [37]) less than or equal to ε, starting from any initial probability distribution μ. The total variation distance between two distributions is equal to half the l_1 norm of $\pi - \mu$ [37]. We define the SA runtime at inverse temperature β, $\tau_{\text{SA}}(\varepsilon, \beta)$, as the mixing time normalized by the Gibbs population of the optimal independent sets of size α. Explicitly, we let (see [37], Eqs. 4.2 and 4.30)

$$\tau_{\text{SA}}(\varepsilon, \beta) = \frac{1}{n\pi_\alpha} \min\left\{t : \max_\mu \sum_{z \in \{0,1\}^n} |\pi_z - P^t \mu_z| \le \varepsilon\right\}, \tag{A4}$$

where μ is the initial distribution, $P = (P_{z,z'})$ is the matrix of Markov chain transition probabilities, and

$$\pi_b = \sum_{z : H_{\text{cost}}(z) = -\delta b} \pi_z \tag{A5}$$

is the Gibbs population of independent sets of size b. $\tau_{\text{SA}}(\varepsilon, \beta)$ represents the time to sample an optimal solution from the Gibbs distribution. The normalization factor of $1/\pi_\alpha$ is necessary because at high temperatures, the mixing time may be small, but the Gibbs population of the optimal solutions is correspondingly very small. At low temperatures, the normalization factor is unnecessary because optimal solutions have high Gibbs population. As a result, the mixing time is directly related to

the time to find an optimal solution, maximized over all optimal solutions (i.e., the *hitting time*) [65]. We define the SA runtime $\tau_{\rm SA}(\varepsilon)$ as the minimum runtime over all temperatures,

$$\tau_{\rm SA}(\varepsilon) = \min_{\beta} \tau_{\rm SA}(\varepsilon, \beta). \tag{A6}$$

Our main result, stated next, is an analytic lower bound on $\tau_{\rm SA}(\varepsilon)$.

Theorem 1. *Consider any Metropolis-Hastings SA algorithm that prepares the Gibbs distribution of the Maximum Independent Set cost Hamiltonian $H_{\rm cost}$. Suppose the SA update rule alters at most k of the n total spins. Define a cutoff independent set size $b^\star$, such that the number of larger independent sets is decreasing, i.e. $D_{b-1}/D_b \geq 1$ for $b > b^\star$. Then for any error $\varepsilon < 1/2$, the SA runtime $\tau_{\rm SA}(\varepsilon)$ can be lower-bounded as*

$$\tau_{\rm SA}(\varepsilon) \geq \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2nk} \max_{b>b^\star} \frac{D_{b-1}}{D_b}. \tag{A7}$$

Before proceeding, we note that the restriction $b > b^\star$ appearing in Theorem 1 is not necessary when the independence polynomial of the graph is *unimodal*, meaning that $D_0 \leq D_1 \leq \cdots \leq D_{b^\star} \geq \cdots \geq D_{\alpha-1} \geq D_\alpha$. This condition is met for every unit-disk graph we study in Appendix B.

Proof. The SA runtime at temperature $1/\beta$ can be lower-bounded by the inverse of the spectral gap $\Delta_{\rm SA} = \Delta_{\rm SA}(\beta)$ between the largest and second largest eigenvalue of the corresponding Markov chain matrix P with transition probabilities $P_{z,z'}$ as ([37], Eq. 12.14)

$$\tau_{\rm SA}(\varepsilon, \beta) \geq \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{n\pi_\alpha} \left(\frac{1}{\Delta_{\rm SA}} - 1\right). \tag{A8}$$

Because $\frac{1}{\Delta_{\rm SA}} \gg 1$, we will ignore the second term. This bound applies to any Markov chain transition matrix P which satisfies detailed balance and is *lazy*, meaning that the outwards transition probability $\sum_{z':z'\neq z} P_{z,z'} \leq 1/2$ for any $|z\rangle$. Any Markov chain P can be made lazy by taking $(P + \mathbb{1})/2$ (i.e., adding weight-1/2 self-loops to each $|z\rangle$). This transformation does not substantially affect the mixing time because it reduces the outwards transition probability by at most a factor of 2, so we will analyze P instead of $(P+\mathbb{1})/2$. Note that in Eq. (A8) we divided the standard definition of mixing time by n because we allow the SA algorithm to "parallelize" updates over different spins.

We can therefore lower-bound $\tau_{\rm SA}(\varepsilon, \beta)$ by upper bounding $\Delta_{\rm SA}$ and π_α. To do this, we use the Cheeger inequality [38], which can be used to establish an upper bound on $\Delta_{\rm SA}$ for any Markov chain satisfying detailed balance. The idea in a Cheeger bound is to bipartition the state space of the Markov chain into two sets, S and S^c, such that in the Gibbs distribution π, very little probability flows from S to S^c during one update of the Markov chain. The spectral gap is then upper bounded by this probability flow Q_{S,S^c} normalized by the total Gibbs population π_S in S. Explicitly, the Cheeger inequality states

$$\Delta_{\rm SA} \leq \frac{2Q_{S,S^c}}{\pi_S}, \qquad \pi_S = \sum_{z\in S} \pi_z, \tag{A9}$$

for any S with $\pi_S < \frac{1}{2}$, where

$$\begin{aligned} Q_{S,S^c} &= \sum_{z\in S, z'\in S^c} \pi_z P_{z,z'} \\ &= \sum_{z\in S, z'\in S^c} \pi_{z'} P_{z',z} \\ &= Q_{S^c,S} \end{aligned} \tag{A10}$$

is the flow from S to S^c. Note that $Q_{S,S^c} = Q_{S^c,S}$ follows from the detailed balance condition on P in Eq. (A3). When Q_{S,S^c} is small, $\Delta_{\rm SA}$ is correspondingly small by Eq. (A9) and the SA runtime is large by Eq. (A8).

We will first consider the low temperature case $\pi_S < 1/2$, and obtain an upper bound on $Q_{S,S^c}/\pi_S$. Let $k \in \{1, 2, \ldots, n\}$ denote the maximum number of spins altered during a proposed update, and $b \in \{b^\star, b^\star + 1, \ldots, \alpha\}$ represent a particular independent set size satisfying $b > b^\star$. We define the set

$$S = \{z : H_{\rm cost}(z) \geq -\delta(b-1)\} \tag{A11}$$

of independent sets of size $b-1$ or smaller. We first replace all the probabilities π_z in Eq. (A10) with $e^{\beta\delta(b-1)}/\mathcal{Z}_\beta$. This gives an upper bound on $Q_{S,S^c}/\pi_S$, because $H_{\rm cost} = -\delta(b-1)$ is the smallest value of $H_{\rm cost}$ present in S:

$$\frac{Q_{S,S^c}}{\pi_S} \leq \frac{e^{\beta\delta(b-1)}}{\pi_S \mathcal{Z}_\beta} \sum_{\substack{H_{\rm cost}(z)\geq -\delta(b-1) \\ H_{\rm cost}(z')\leq -\delta b}} P_{z,z'}. \tag{A12}$$

Now, plugging in the Metropolis-Hastings update rule from Eq. (A2), we have

$$\begin{aligned} \frac{Q_{S,S^c}}{\pi_S} &\leq \frac{e^{\beta\delta(b-1)}}{\pi_S \mathcal{Z}_\beta} \sum_{\substack{z':H_{\rm cost}(z')\leq -\delta b \\ z:H_{\rm cost}(z)\geq -\delta(b-1)}} p_{z,z'} \min\left(1, \frac{e^{-\beta[H_{\rm cost}(z')]}}{e^{-\beta H_{\rm cost}(z)]}}\right) \\ &= \frac{e^{\beta\delta(b-1)}}{\pi_S \mathcal{Z}_\beta} \sum_{H_{\rm cost}(z')\leq -\delta b} \left(\sum_{H_{\rm cost}(z)\geq -\delta(b-1)} p_{z',z} \right). \end{aligned} \tag{A13}$$

where in the second line we have used that $p_{z,z'} = p_{z',z}$ under the detailed balance condition [Eq. (A3)]. The inner summation over configurations $|z\rangle$ at fixed $|z'\rangle$ is equal to the probability of proposing an update from $|z'\rangle$ to any configuration $|z\rangle$ with $H_{\rm cost}(z) \geq -\delta(b-1)$. This probability is at most one because the total transition probability out of $|z'\rangle$ into S is at most one, and is strictly

zero if $H_{\text{cost}}(z') < -\delta(b+k-1)$ (because we have assumed that we update at most k spins). This constraint yields

$$\begin{aligned}\frac{Q_{S,S^c}}{\pi_S} &\le \frac{e^{\beta\delta(b-1)}}{\pi_S \mathcal{Z}_\beta} \sum_{-\delta\min(\alpha,b+k-1)\le H_{\text{cost}}(z')\le -\delta b} 1 \\ &= \frac{e^{\beta\delta(b-1)}}{\pi_S \mathcal{Z}_\beta} \sum_{b'=b}^{\min(\alpha,b+k-1)} D_{b'} \\ &\le \frac{k D_b e^{\beta\delta(b-1)}}{\pi_S \mathcal{Z}_\beta} \\ &\le \frac{k D_b}{D_{b-1}}. \end{aligned} \tag{A14}$$

In the third step we used the fact that $D_b \ge D_{b'}$ for any $b' > b^\star$, and in the fourth step we have used $\pi_S = \sum_{b'=0}^{b-1} D_{b'} e^{\beta\delta b'}/\mathcal{Z}_\beta > D_{b-1} e^{\beta\delta(b-1)}/\mathcal{Z}_\beta$. From Eq. (A9), the SA spectral gap Δ_{SA} is thus bounded as

$$\Delta_{\text{SA}} \le \frac{2Q_{S,S^c}}{\pi_S} \le \frac{2kD_b}{D_{b-1}}. \tag{A15}$$

Combining this with the lower bound on runtime $\tau_{\text{SA}}(\varepsilon,\beta)$ [Eq. (A8)], and plugging in $\pi_\alpha \le 1$, we have for any β such that $\pi_S < 1/2$,

$$\tau_{\text{SA}}(\varepsilon,\beta) \ge \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2nk} \frac{D_{b-1}}{D_b}. \tag{A16}$$

On the other hand, at high temperatures $\pi_S > 1/2$, we must swap S with S^c in the Cheeger bound [Eq. (A9)],

$$\Delta_{\text{SA}}\pi_\alpha \le \frac{2Q_{S^c,S}\pi_\alpha}{\pi_{S^c}} = \frac{2Q_{S,S^c}\pi_\alpha}{\pi_{S^c}} \tag{A17}$$

where we have used the fact that $Q_{S,S^c} = Q_{S^c,S}$. By Eq. (A14) we have $Q_{S,S^c} \le kD_b e^{\beta\delta(b-1)}/\mathcal{Z}_\beta$, so we find

$$\Delta_{\text{SA}}\pi_\alpha \le \frac{2kD_b e^{\beta\delta(b-1)}\pi_\alpha}{\mathcal{Z}_\beta \pi_{S^c}} \le \frac{2kD_b e^{\beta\delta(b-1)}}{\mathcal{Z}_\beta}, \tag{A18}$$

using $\pi_\alpha \le \pi_{S^c}$ (because sets of size α are contained in S^c). Now, since $\mathcal{Z}_\beta > D_{b-1}e^{\beta\delta(b-1)}$, we are left with

$$\Delta_{\text{SA}}\pi_\alpha \le \frac{2kD_b}{D_{b-1}}, \tag{A19}$$

which gives the same bound as in the low-temperature case via Eq. (A8). Because the same bound holds for all temperatures and for any $b > b_\star$, we can use Eq. (A6) to obtain a lower bound on $\tau_{\text{SA}}(\varepsilon)$, which gives us Theorem 1.

Finally, we note that Theorem 1 can be applied to general combinatorial optimization problems with discrete cost Hamiltonian energies. Our proof does not change if we replace the energies of H_{cost}, $\{-\delta b\}_{b=0,1,\ldots,\alpha}$, with energies $\{E_b\}_{b=0,1,\ldots,\alpha}$ for any generic cost function with $\alpha+1$ discrete energy levels, and let D_b represent the number of spin configurations with energy E_b. As a result, Theorem 1 can be applied to generic discrete cost functions beyond Maximum Independent Set.

2. Parallel tempering

We now derive a runtime lower bound for a wide class of parallel tempering algorithms using the Metropolis-Hastings update rule. Because our bound uses identical techniques to the runtime lower bound for SA, we recommend the reader read Appendix A 1 before proceeding. In parallel tempering there are M copies, or *replicas*, of the n-spin system of SA, each equilibrating to the Gibbs distribution of H_{cost} at temperatures $1/\beta_1,\ldots,1/\beta_M$. The state space is the product of states over all the replicas $\{z_1\ldots z_M\}$, where $z_i \in \{0,1\}^n$ represents the spin configuration of the ith replica. Similar to SA, the state of a single replica can be updated based on proposing an update to at most k spins. However, in parallel tempering collective updates involving multiple replicas are also possible. We will consider collective Metropolis-Hastings update rules,

$$\begin{aligned}&P_{z_1\ldots z_M, z_1'\ldots z_M'} \\ &\quad = p_{z_1\ldots z_M, z_1'\ldots z_M'} \min\left(1, e^{-\sum_{i=1}^M \beta_i[H_{\text{cost}}(z_i')-H_{\text{cost}}(z_i)]}\right),\end{aligned} \tag{A20}$$

where $p_{z_1\ldots z_M, z_1'\ldots z_M'}$ is the probability of proposing an update to the configuration $z_1'\ldots z_M'$ given that the current configuration is $z_1\ldots z_M$. Note that this update rule satisfies the detailed balance condition in Eq. (A3). The equilibrium distribution is therefore the Gibbs distribution,

$$\pi_{z_1\ldots z_M} = \frac{e^{-\sum_{i=1}^M \beta_i H_{\text{cost}}(z_i)}}{\prod_{i=1}^M \mathcal{Z}_{\beta_i,i}}, \quad \mathcal{Z}_{\beta_i,i} = \sum_{b=0}^{\alpha} D_b e^{\beta_i \delta b}. \tag{A21}$$

We define the parallel tempering runtime $\tau_{\text{PT}}(\varepsilon)$ as

$$\tau_{\text{PT}}(\varepsilon) = \min_{\beta_1\ldots\beta_M} \tau_{\text{PT}}(\varepsilon,\beta_1\ldots\beta_M), \tag{A22}$$

where $\tau_{\text{PT}}(\varepsilon,\beta_1\ldots\beta_M)$ is the runtime lower bound for replica temperatures $\beta_1\ldots\beta_M$ defined similarly to SA [Eq. (A4)]:

$$\begin{aligned}&\tau_{\text{PT}}(\varepsilon,\beta_1\ldots\beta_M) \\ &\quad = \frac{M}{n\pi_\alpha} \min\left\{t : \max_{\mu} \sum_{z\in\{0,1\}^n} |\pi_z - P^t\mu_z| \le \varepsilon\right\},\end{aligned} \tag{A23}$$

where P is the parallel tempering Markov chain, μ is the initial probability distribution, and

$$\pi_\alpha = \sum_{i=1}^M \sum_{\substack{z_1\ldots z_M: \\ H_{\text{cost}}(z_i)=-\delta b}} \pi_{z_1\ldots z_M} \tag{A24}$$

is now the probability that the configuration of at least one replica is an independent set of size b. Note that $\tau_{\text{PT}}(\varepsilon,\beta_1\ldots\beta_M)$ in Eq. (A23) has a factor of M in the numerator. This is because we allow the parallel tempering update rule to update the spin configuration on

all M replicas; thus, the time complexity to perform an update is $\mathcal{O}(M)$. This also excludes the possibility of a trivial "speedup" from making M exponentially large, at the expense of, e.g., $M = \mathcal{O}(2^n)$ space-time complexity.

a. Replica exchange, arbitrary single-replica updates, and constant-sized collective updates

We first consider parallel tempering algorithms that include the following update rules: single-replica updates that can update an arbitrary number of spins k on a single replica, collective-replica updates that modify k' spins on each replica, where k' is restricted to be constant in n, and replica exchange updates. Replica exchange updates are defined as proposing to exchange the states z_i and z_j of two replicas i and j. Our runtime lower bound is stated next in Theorem 2. We will generalize our result to include non-local *isoenergetic cluster updates* [66] later in Theorem 3.

Theorem 2. *Consider a parallel tempering algorithm with M replicas and any update rule as described above. Define a cutoff independent set size $b^\star$, such that the number of larger independent sets is decreasing, i.e. $D_{b-1}/D_b \geq 1$ for $b > b^\star$. Then for any error $\varepsilon < 1/2$, the parallel tempering runtime $\tau_{\mathrm{PT}}(\varepsilon)$ is bounded as*

$$\tau_{\mathrm{PT}}(\varepsilon) \geq \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2nk'n^{k'}} \max_{b>b_\star} \frac{D_{b-1}}{D_b}. \tag{A25}$$

Proof. Define the set $\mathcal{S}$ as the set of states with all the replicas having independent set size less than b, for $b > b_\star$,

$$\begin{aligned}\mathcal{S} &= \{z_1 \dots z_M : \forall i \in \{1,\dots,M\}, H_{\mathrm{cost}}(z_i) \geq -\delta(b-1)\} \\ &= S_1 \times \cdots \times S_M,\end{aligned} \tag{A26}$$

where S_i is the partition defined for a single replica as defined in Eq. (A11). As with the SA runtime lower bound in Appendix A 1, our goal is to bound the flow of probability $Q_{\mathcal{S},\mathcal{S}^c}$ from $\mathcal{S}$ to $\mathcal{S}^c$ in the Gibbs distribution,

$$Q_{\mathcal{S},\mathcal{S}^c} = \sum_{\substack{z_1\dots z_M \in \mathcal{S} \\ z'_1 \dots z'_M \in \mathcal{S}^c}} \pi_{z_1\dots z_M} P_{z_1\dots z_M, z'_1\dots z'_M}, \tag{A27}$$

to obtain a Cheeger bound on the spectral gap of the parallel tempering Markov chain $\Delta_{\mathrm{PT}} = \Delta_{\mathrm{PT}}(\beta_1 \dots \beta_M)$,

$$\Delta_{\mathrm{PT}} \leq \frac{2Q_{\mathcal{S},\mathcal{S}^c}}{\pi_{\mathcal{S}}} = \frac{2\sum_{\substack{z_1\dots z_M \in \mathcal{S} \\ z'_1\dots z'_M \in \mathcal{S}^c}} \pi_{z_1\dots z_M} P_{z_1\dots z_M, z'_1\dots z'_M}}{\prod_{i=1}^M \pi_{S_i}} \tag{A28}$$

where π_{S_i} is the Gibbs population of S_i on a replica i, as in Eq. (A9), and $\pi_{\mathcal{S}}$ is the Gibbs population of $\mathcal{S}$. Eq. (A28) gives a lower bound on the runtime via Eq. (A8). From Eq. (A28) we can immediately see that replica exchange updates do not contribute to $Q_{\mathcal{S},\mathcal{S}^c}$ because swapping the states of two replicas in $\mathcal{S}$ does not transfer probability from $\mathcal{S}$ to $\mathcal{S}^c$. In addition, arbitrary updates to a single replica are subject to the same bound as SA [Eq. (A7)]. Therefore, it only remains to bound collective updates that update at most k' spins on each replica, where k' is constant in n. The runtime lower bound is then given by the minimum of the runtime lower bounds on collective updates and single-replica updates. We will find that the runtime lower bound for collective updates is smaller than for single-replica updates; hence, Theorem 2 reflects the collective update bound.

As before, we will obtain an upper bound on $Q_{\mathcal{S},\mathcal{S}^c}$. Notice that only transitions from configurations in $\mathcal{S}$ within k' spin flips of some $z_1 \dots z_M \in \mathcal{S}^c$ can contribute to $Q_{\mathcal{S},\mathcal{S}^c}$. We denote these configurations as $\partial\mathcal{S}$. Configurations in $\partial\mathcal{S}$ must have least one replica j within k' spin flips of S_j^c, whereas all other replicas i may be in any configuration in S_i. We let ∂S_j denote configurations $z_j \in S_j$ within k' spin flips of S_j^c, and $\pi_{\partial S_j} = \sum_{z_j \in \partial S_j} \pi_{z_j}$. We then can show

$$\begin{aligned} Q_{\mathcal{S},\mathcal{S}^c} &\leq \sum_{z_1\dots z_M \in \partial\mathcal{S}} \pi_{z_1\dots z_M} \\ &\leq \sum_{j=1}^M \pi_{\partial S_j} \prod_{\substack{i=1 \\ i\neq j}}^M \pi_{S_i} = \sum_{j=1}^M \pi_{\partial S_j} \frac{\pi_{\mathcal{S}}}{\pi_{S_j}} \\ &\leq (k')^2 \binom{n}{k'} \sum_{j=1}^M \frac{D_b e^{\delta\beta_j(b-1)}}{\mathcal{Z}_{\beta_j,j}} \frac{\pi_{\mathcal{S}}}{\pi_{S_j}} \\ &\leq k' n^{k'} \sum_{j=1}^M \frac{D_b e^{\delta\beta_j(b-1)}}{\mathcal{Z}_{\beta_j,j}} \frac{\pi_{\mathcal{S}}}{\pi_{S_j}}, \end{aligned} \tag{A29}$$

where in the third line we have used the fact that the Gibbs population of any configuration in ∂S_j is $\leq e^{\delta\beta_j(b-1)}/\mathcal{Z}_{\beta_j,j}$, and the fact that there are $\leq (k')^2\binom{n}{k'} D_b$ such configurations. Then we may write

$$\begin{aligned} \frac{Q_{\mathcal{S},\mathcal{S}^c}}{\pi_{\mathcal{S}}} &\leq k' n^{k'} \sum_{j=1}^M \frac{D_b e^{\delta\beta_j(b-1)}}{\mathcal{Z}_{\beta_j,j}\pi_{S_j}} \\ &\leq k' n^{k'} \sum_{j=1}^M \frac{D_b}{D_{b-1}} = k' n^{k'} M \frac{D_b}{D_{b-1}}, \end{aligned} \tag{A30}$$

where in the second line we have used that $\pi_{S_j} \geq D_{b-1} e^{\beta_j\delta(b-1)}/\mathcal{Z}_{\beta_j,j}$. We can combine Eqs. (A30) and (A9) and use the fact that $\pi_\alpha \leq 1$ to get

$$\Delta_{\mathrm{PT}} \leq 2k' n^{k'} M \min_{b>b_\star} \frac{D_b}{D_{b-1}} \tag{A31}$$

for the spectral gap of the parallel tempering Markov chain Δ_{PT}. This, combined with Eq. (A8), yields a runtime lower bound for parallel tempering given by Eq. (A25). We emphasize that this bound is only restrictive when $\max_{b>b_\star}(D_b/D_{b-1})$ is much larger than $n^{k'}$.

Because $\max_{b>b_\star}(D_b/D_{b-1})$ grows exponentially in $\sqrt{n}$ in the worse case for the Maximum Independent Set problem on unit-disk graphs (see Appendix B), this bound is only useful when k' does not grow with n.

The above result holds when $\pi_{\mathcal{S}} < 1/2$. Just as in the case of SA, we can derive the same bound on the runtime when $\pi_{\mathcal{S}} > 1/2$. Using the fact that $Q_{\mathcal{S},\mathcal{S}^c} = Q_{\mathcal{S}^c,\mathcal{S}}$ [see Eq. (A10)], we use Eq. (A29) to receive:

$$\begin{aligned}\Delta_{\mathrm{PT}}\pi_\alpha &\leq \frac{2Q_{\mathcal{S},\mathcal{S}^c}\pi_\alpha}{\pi_{\mathcal{S}^c}} \leq 2Q_{\mathcal{S},\mathcal{S}^c} \\ &\leq 2k'n^{k'}M \min_{b>b_\star}\frac{D_b}{D_{b-1}}. \end{aligned} \tag{A32}$$

As a result, the same bound Eq. (A25) holds for the case where $\pi_{\mathcal{S}} > 1/2$.

Finally, we note that Theorem 2 can be applied to general combinatorial optimization problems with discrete cost Hamiltonian energies. Our proof, as in the case of SA, does not change if we replace the energies of H_{cost}, $\{-\delta b\}_{b=0,1,\ldots,\alpha}$, with energies $\{E_b\}_{b=0,1,\ldots,\alpha}$ for any generic cost function with $\alpha+1$ discrete energy levels, and let D_b represent the number of spin configurations with energy E_b.

b. *Isoenergetic cluster updates*

Here we obtain a runtime lower bound for all parallel tempering algorithms that use the same update rules as in the previous Appendix A 2, in addition to *isoenergetic cluster updates*, which are non-local updates designed specifically for optimizing two-dimensional spin glasses [66]. Isoenergetic cluster updates collectively update a pair of replicas i, j by identifying clusters of spins (vertices) connected by edges for which z_i and z_j differ. The update rule then proposes to exchange the configurations of spins within a randomly chosen connected cluster between z_i and z_j. One can check that this update rule conserves the total energy of the two replicas: $H_{\mathrm{cost}}(z_i) + H_{\mathrm{cost}}(z_j) = H_{\mathrm{cost}}(z_i') + H_{\mathrm{cost}}(z_j')$, where z_i' and z_j' are the spin configurations after an isoenergetic cluster update. Note that isoenergetic cluster updates are equivalent to replica exchange updates when there is only a single connected cluster of differing spins.

The bound that we will derive in Theorem 3 is similar to the parallel tempering runtime lower bound previously derived in Theorem 2 when $\min_{b>b^\star}(D_b/D_{b-1})$ is small compared to the other ratios D_b/D_{b-1}, i.e. when there is a single smallest coupling that limits the runtime. In Fig. 6 we numerically find that the scaling of our bound, stated next in Theorem 3, is similar to the SA runtime lower bound in Theorem 1 for the top 5% hardest instances of each system size studied in Appendix B. In particular, Fig. 6 plots $\max_{b>b^\star}\left[D_b/D_{b-1} + \sum_{b_1'=b}^{\alpha}\sum_{b_2'=0}^{2(b-1)-b_1}\sum_{k=b_1'-b+1}^{b-1-b_2'}(D_{b_1'}D_{b_2'})(D_{b_1'-k}D_{b_2'+k})\right]^{-1}$ versus the quantity $\max_{b>b^\star}(D_{b-1}/D_b)$. These quantities are equal to the parallel tempering and SA runtime lower bounds in Theorems 1 and 3, respectively, up to subleading polynomial factors in n.

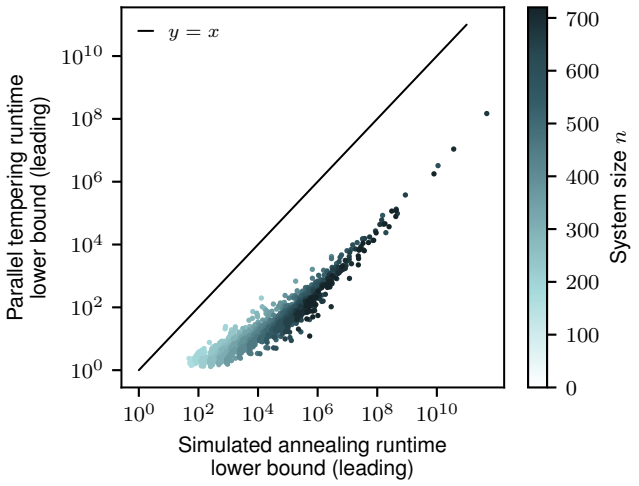

FIG. 6. Simulated annealing and parallel tempering runtime lower bounds. We plot $\max_{b>b^\star}\left[D_b/D_{b-1} + \sum_{b_1'=b}^{\alpha}\sum_{b_2'=0}^{2(b-1)-b_1}\sum_{k=b_1'-b+1}^{b-1-b_2'}(D_{b_1'}D_{b_2'})(D_{b_1'-k}D_{b_2'+k})\right]^{-1}$ versus $\max_{b>b^\star}(D_{b-1}/D_b)$ for the top 5% hardest instances of each system size studied in Appendix B. These quantities are equal to the SA and parallel tempering runtime lower bounds in Theorems 1 and 3, respectively, up to subleading polynomial factors in n.

Theorem 3. *Consider a parallel tempering algorithm with M replicas using isoenergetic cluster updates as described above, in combination with the updates described in Theorem 2. Then for any error $\varepsilon < 1/2$, the parallel tempering runtime $\tau_{\mathrm{PT}}(\varepsilon)$ is bounded as*

$$\begin{aligned}\tau_{\mathrm{PT}}(\varepsilon) \geq \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2n}\max_{b>b^\star}\Bigg[k'n^{k'}\frac{D_b}{D_{b-1}} \\ + \sum_{b_1'=b}^{\alpha}\sum_{b_2'=0}^{2(b-1)-b_1}\sum_{k=b_1'-b+1}^{b-1-b_2'}\frac{D_{b_1'}D_{b_2'}}{D_{b_1'-k}D_{b_2'+k}}\Bigg]^{-1}.\end{aligned} \tag{A33}$$

Proof. As before, we bound the flow of probability $Q_{\mathcal{S},\mathcal{S}^c}$ from $\mathcal{S}$ to $\mathcal{S}^c$ in the Gibbs distribution,

$$Q_{\mathcal{S},\mathcal{S}^c} = \sum_{\substack{z_1\ldots z_M\in\mathcal{S}\\ z_1'\ldots z_M'\in\mathcal{S}^c}} \pi_{z_1\ldots z_M}P_{z_1\ldots z_M,z_1'\ldots z_M'}, \tag{A34}$$

We define $\mathcal{S}$ identically to Eq. (A26). As a result, the runtime lower bound we will derive in Theorem 3 automatically applies to the same update rules from Theorem 2, and it only remains to upper bound $Q_{\mathcal{S},\mathcal{S}^c}$ for isoenergetic cluster updates. The total $Q_{\mathcal{S},\mathcal{S}^c}$ will then be bounded by the sum of the bounds on $Q_{\mathcal{S},\mathcal{S}^c}$ derived here for isoenergetic cluster updates and on the bound in

Theorem 2. The inverse of this sum of bounds will yield the bound in Theorem 3.

An isoenergetic cluster update first proposed to update the configurations of a pair of replicas, which are chosen according some probability distribution. Without loss of generality, we will call these replicas 1 and 2, and denote the probability they are proposed as p_{12}. Once a pair of replicas is proposed, the quantity $Q_{\mathcal{S},\mathcal{S}^c}/\pi_{\mathcal{S}}$ is independent of the remaining replicas. Thus, we may consider the flow $Q^{(12)}_{\mathcal{S},\mathcal{S}^c}$ on only replicas 1 and 2. We may bound the flow as

$$\begin{aligned} Q^{(12)}_{\mathcal{S},\mathcal{S}^c} &= p_{12} \sum_{\substack{z_1 z_2 \in \mathcal{S} \\ z_1' z_2' \in \mathcal{S}^c}} \pi_{z_1' z_2'} P_{z_1' z_2', z_1 z_2} \\ &= p_{12} \sum_{\substack{z_1 z_2 \in \mathcal{S} \\ z_1' z_2' \in \mathcal{S}^c}} \pi_{z_1' z_2'} p_{z_1' z_2', z_1 z_2} \min\left(1, \frac{\pi_{z_1 z_2}}{\pi_{z_1' z_2'}}\right) \\ &\le p_{12} \sum_{\substack{z_1 z_2 \in \mathcal{S} \\ z_1' z_2' \in \mathcal{S}^c}} \pi_{z_1 z_2} p_{z_1' z_2', z_1 z_2}, \end{aligned} \tag{A35}$$

where $P_{z_1' z_2', z_1 z_2}$ is the probability of updating to $z_1' z_2'$ given that the current configuration of the two replicas we have chosen to update is $z_1 z_2$. Since $z_1' z_2' \in \mathcal{S}^c$, at least one of z_1' or z_2' must be in S_1^c or S_2^c. We assume without loss of generality that it is z_1', so that $H_{\text{cost}}(z_1') \le -\delta b$. Then, if z_1', z_2' can isoenergetically update to $z_1, z_2 \in \mathcal{S}$, we must have $H_{\text{cost}}(z_2') \ge -2\delta(b-1) - H_{\text{cost}}(z_1')$, because the combined energy of z_1', z_2' must be at least $-2\delta(b-1)$. Furthermore, the number of spins that can be exchanged between the two replicas is lower-bounded by the restriction that $z_1 \in S_1$ and upper-bounded by the restriction that $z_2 \in S_2$. As a result, the sum can be parameterized as

$$\begin{aligned} Q^{(12)}_{\mathcal{S},\mathcal{S}^c} &\le p_{12} \sum_{\substack{z_1 z_2 \in \mathcal{S} \\ z_1' z_2' \in \mathcal{S}^c}} \pi_{z_1 z_2} p_{z_1' z_2', z_1 z_2} \qquad \text{(A36)} \\ &\le p_{12} \sum_{b_1'=b}^{\alpha} \sum_{b_2'=0}^{2(b-1)-b_1} \sum_{k=b_1'-b+1}^{b-1-b_2'} \sum_{\substack{H_{\text{cost}}(z_1')=-\delta b_1' \\ H_{\text{cost}}(z_2')=-\delta b_2' \\ H_{\text{cost}}(z_1)=-\delta(b_1'-k) \\ H_{\text{cost}}(z_2)=-\delta(b_2'+k)}} \pi_{z_1 z_2} p_{z_1' z_2', z_1 z_2} \\ &\le p_{12} \sum_{b_1'=b}^{\alpha} \sum_{b_2'=0}^{2(b-1)-b_1} \sum_{k=b_1'-b+1}^{b-1-b_2'} D_{b_1'} D_{b_2'} \frac{e^{\beta_1\delta(b_1'-k)+\beta_2\delta(b_2'+k)}}{\mathcal{Z}_{\beta_1,1}\mathcal{Z}_{\beta_2,2}}. \end{aligned}$$

In the third line, we used the facts that $\pi_{z_1 z_2} = e^{\beta_1\delta(b_1'-k)+\beta_2\delta(b_2'+k)}/(\mathcal{Z}_{\beta_1,1}\mathcal{Z}_{\beta_2,2})$ and $\sum_{z_1 z_2} p_{z_1' z_2', z_1 z_2} \le 1$, then replaced $\sum_{z_1' z_2'}$ with $D_{b_1'} D_{b_2'}$. To remove the factors of β_1 and β_2, we may also use the fact that $\mathcal{Z}_{\beta_1,1}$ contains a $D_{b_1'-k} e^{\beta_1\delta(b_1'-k)}$ term and $\mathcal{Z}_{\beta_2,2}$ contains a $D_{b_2'+k} e^{\beta_2\delta(b_2'+k)}$ term, to obtain

$$Q^{(12)}_{\mathcal{S},\mathcal{S}^c} \le p_{12} \sum_{b_1'=b}^{\alpha} \sum_{b_2'=0}^{2(b-1)-b_1} \sum_{k=b_1'-b+1}^{b-1-b_2'} \frac{D_{b_1'} D_{b_2'}}{D_{b_1'-k} D_{b_2'+k}}. \tag{A37}$$

Now summing over all replicas (not just 1, 2) that could be proposed for replica updates and using $\sum_{ij} p_{ij} \le 1$, we arrive at

$$Q_{\mathcal{S},\mathcal{S}^c} \le \sum_{b_1'=b}^{\alpha} \sum_{b_2'=0}^{2(b-1)-b_1} \sum_{k=b_1'-b+1}^{b-1-b_2'} \frac{D_{b_1'} D_{b_2'}}{D_{b_1'-k} D_{b_2'+k}}. \tag{A38}$$

For reasons analogous to those given in Appendix A 1, this is sufficient to establish the bound in Theorem 3 when $\pi_{\mathcal{S}} > 1/2$. When $\pi_{\mathcal{S}} < 1/2$, we instead revert to Eq. (A36) and compute the bound as

$$\begin{aligned} \frac{Q^{(12)}_{\mathcal{S},\mathcal{S}^c}}{\pi_{\mathcal{S}}} &\le p_{12} \pi_{\mathcal{S}}^{-1} \\ &\times \sum_{b_1'=b}^{\alpha} \sum_{b_2'=0}^{2(b-1)-b_1} \sum_{k=b_1'-b+1}^{b-1-b_2'} D_{b_1'} D_{b_2'} \frac{e^{\beta_1\delta(b_1'-k)+\beta_2\delta(b_2'+k)}}{\mathcal{Z}_1 \mathcal{Z}_2} \\ &\le p_{12} \Big(\sum_{b_1=0}^{b} D_{b_1} e^{\beta_1\delta b_1}\Big)^{-1} \Big(\sum_{b_2=0}^{b_1} D_{b_2} e^{\beta_2\delta b_2}\Big)^{-1} \\ &\times \sum_{b_1'=b}^{\alpha} \sum_{b_2'=0}^{2(b-1)-b_1} \sum_{k=b_1'-b+1}^{b-1-b_2'} D_{b_1'} D_{b_2'} e^{\beta_1\delta(b_1'-k)+\beta_2\delta(b_2'+k)}. \end{aligned} \tag{A39}$$

At this point, we again use the fact that for every (b_1', b_2', k) term in the numerator, the denominator contains a term $D_{b_1'-k} e^{\beta_1\delta(b_1'-k)} D_{b_2'+k} e^{\beta_2\delta(b_2'+k)}$, allowing us to arrive at

$$\frac{Q^{(12)}_{\mathcal{S},\mathcal{S}^c}}{\pi_{\mathcal{S}}} \le p_{12} \sum_{b_1'=b}^{\alpha} \sum_{b_2'=0}^{2(b-1)-b_1} \sum_{k=b_1'-b+1}^{b-1-b_2'} \frac{D_{b_1'} D_{b_2'}}{D_{b_1'-k} D_{b_2'+k}}, \tag{A40}$$

from which we can establish the bound in Theorem 3 after summing over all choices of 1, 2.

3. Quantum Monte Carlo

We now establish a runtime lower bound for a wide class of QMC algorithms. Our bound uses identical techniques to the analytic runtime lower bounds of SA (Appendix A 1) and parallel tempering (Appendix A 2), which we recommend the reader read first for context. We consider path-integral QMC algorithms which are designed to sample from the populations of the Gibbs distribution of the modified QAA Hamiltonian,

$$\rho_{zz} = \frac{\langle z|\, e^{-\beta(H_{\text{QAA}}+\lambda H_\ell)}\, |z\rangle}{\mathcal{Z}_\beta}, \quad \mathcal{Z}_\beta = \text{Tr}\big(e^{-\beta(H_{\text{QAA}}+\lambda H_\ell)}\big). \tag{A41}$$

We can write the partition function $\mathcal{Z}_\beta$ in the z-basis by Trotterizing $H = H_{\text{QAA}} + \lambda H_\ell$ and inserting copies of the

identity matrix. Although we do not assume a particular form of Trotterization of $\mathcal{Z}_\beta$, we may take, for example,

$$\begin{aligned}\mathcal{Z}_\beta &= \sum_{z_1} \langle z_1| \, e^{-\beta(H_{\mathrm{QAA}}+\lambda H_\ell)} \, |z_1\rangle \\ &\simeq \sum_{z_1} \langle z_1| \left(e^{-\beta H_{\mathrm{o.d.}}/M} e^{-\beta H_{\mathrm{d}}/M}\right)^M |z_1\rangle \\ &= \sum_{z_1 \dots z_M} \langle z_1| \, e^{-\beta H_{\mathrm{o.d.}}/M} \, |z_2\rangle \, \langle z_2| \, e^{-\beta H_{\mathrm{d}}/M} \, |z_2\rangle \\ &\times \langle z_2| \dots |z_M\rangle \, \langle z_M| \, e^{-\beta H_{\mathrm{o.d.}}/M} \, |z_1\rangle \, \langle z_1| \, e^{-\beta H_{\mathrm{d}}/M} \, |z_1\rangle \,, \end{aligned} \tag{A42}$$

where $H_{\mathrm{o.d.}}$ contains only off-diagonal terms of H in the computational basis, and H_{d} contains only diagonal terms in the computational basis. When the number of Trotter steps M is sufficiently large, the marginal probability of configuration $|z_1\rangle$ approximates its population in the Gibbs distribution,

$$\begin{aligned}\pi_{z_1} &= \sum_{z_2 \dots z_M} \pi_{z_1 \dots z_M} \\ &= \rho_{z_1 z_1} \text{ as } M \to \infty, \end{aligned} \tag{A43}$$

where

$$\begin{aligned}\pi_{z_1 \dots z_M} &= \frac{1}{\mathcal{Z}_\beta} \langle z_1| \, e^{-\beta H_{\mathrm{o.d.}}/M} \, |z_2\rangle \, \langle z_2| \, e^{-\beta H_{\mathrm{d}}/M} \, |z_2\rangle \\ &\times \langle z_2| \dots |z_M\rangle \, \langle z_M| \, e^{-\beta H_{\mathrm{o.d.}}/M} \, |z_1\rangle \, \langle z_1| \, e^{-\beta H_{\mathrm{d}}/M} \, |z_1\rangle \end{aligned} \tag{A44}$$

under the particular Trotterization in Eq. (A42). Since the number of Trotter steps needed to obtain a good approximation of $\mathcal{Z}_\beta$ is typically polynomial in β and the norm of H, we consider finite but large $U \gg |\delta|, \beta$.

Path-integral QMC can be used to sample configurations from the distribution $\pi_{z_1 \dots z_M}$. The Metropolis-Hastings update rule updates configuration $z_1 \dots z_M$ to $z'_1 \dots z'_M$ with probability

$$P_{z_1 \dots z_M, z'_1 \dots z'_M} = p_{z_1 \dots z_M, z'_1 \dots z'_M} \min\left(1, \frac{\pi_{z'_1 \dots z'_M}}{\pi_{z_1 \dots z_M}}\right), \tag{A45}$$

where $p_{z_1 \dots z_M, z'_1 \dots z'_M}$ is the probability of proposing an update to $z'_1 \dots z'_M$ given that the current configuration is $z_1 \dots z_M$.

We define the QMC runtime analogously to parallel tempering, as

$$\tau_{\mathrm{QMC}}(\varepsilon) = \min_\beta \tau_{\mathrm{QMC}}(\varepsilon), \tag{A46}$$

where $\tau_{\mathrm{QMC}}(\varepsilon, \beta)$ is the runtime lower bound for QMC at temperature $1/\beta$,

$$\tau_{\mathrm{QMC}}(\varepsilon, \beta) = \frac{M}{n \pi_\alpha} \min\left\{t : \max_\mu \sum_{z \in \{0,1\}^n} |\pi_z - P^t \mu_z| \le \varepsilon\right\}. \tag{A47}$$

where P is the QMC Markov chain, μ is the initial probability distribution, and

$$\pi_\alpha = \sum_{i=1}^{M} \sum_{\substack{z_1 \dots z_M: \\ H_{\mathrm{cost}}(z_i) \le -\delta b}} \pi_{z_1 \dots z_M} \tag{A48}$$

is now the probability that the configuration of at least one replica is an independent set of size b. As with the definition of parallel tempering runtime in Eq. (A23), we include a factor of M in the numerator of Eq. (A47). This decision is justified because we allow the update rule to alter all M Trotter slices, which takes $\mathcal{O}(M)$ time complexity. It also excludes the trivial "speedup" that one might obtain by using exponentially many time slices to enumerate an exponential number of low-energy configurations, at the expense of exponential space complexity. The inclusion of this factor makes our runtime lower bound, stated next in Theorem 4, independent of the parameter M. We will remark in our proof of Theorem 4 that if the number of Trotter slices modified in a single update is $m < M$, then m can be substituted for M in our definition of $\tau_{\mathrm{QMC}}(\varepsilon, \beta)$ in Eq. (A47).

Theorem 4. *Consider any path-integral QMC algorithm which uses a Metropolis-Hastings update rule to modify at most k spins on each of M imaginary time slices, where k is a constant in n. For a given b, let $H^{(r,b)}$ denote the modified QAA Hamiltonian $H = H_{\mathrm{QAA}} + \lambda H_\ell$ restricted to the space of configurations z with $H_{\mathrm{cost}}(z) > -\delta b$, and let $\pi^{(r,b)}$ be the QMC equilibrium distribution associated with $H^{(r,b)}$ at inverse temperature β. Let $|z_{\max}\rangle$ denote the configuration within k spin flips of an independent set $|z\rangle$ with $H_{\mathrm{cost}}(z) = -\delta b$ with the maximum Gibbs population $\pi^{(r,b)}_{z_{\max}}$, and let $e^{(r,b)}_{\max} = \pi^{(r,b)}_{z_{\max}} D_{b-1}$ describe relative enhancement or suppression of its population compared to the uniform superposition state $|S_{b-1}\rangle$. Then the QMC runtime $\tau_{\mathrm{QMC}}(\varepsilon)$ for any error $\varepsilon < 1/2$ is bounded as*

$$\tau_{\mathrm{QMC}}(\varepsilon) \ge \frac{\ln\left(\frac{1}{2\varepsilon}\right)}{2nkn^k} \max_{b > b_\star} \frac{D_{b-1}}{e^{(r,b)}_{\max} D_b}. \tag{A49}$$

We first comment on the implications of Theorem 4 before proceeding to its proof. Denote the Gibbs distribution of $H^{(r,b)}$ as $\rho^{(r,b)} = e^{-\beta H^{(r,b)}}/\mathrm{Tr}(e^{-\beta H^{(r,b)}})$. For the purpose of discussion, assume that M is large enough such that $\pi^{(r,b)}_z$ is a good approximation for $\rho^{(r,b)}_{zz}$. Now, when λ is large enough to ensure that $\rho^{(r,b)}$ is delocalized in the manifold of independent sets of size $b-1$, we have $e^{(r,b)}_{\max} \le 1$. Thus, Eq. (A49) recovers the parallel tempering runtime bound in Theorem 2. Additionally, the QMC runtime is quadratically larger than the QAA runtime. Conversely, if $\rho^{(r,b)}$ is favorably localized among sets of size $\le b-1$ within Hamming distance k of sets of size b, Eq. (A49) yields a weak bound. In particular, if $e^{(r,b)}_{\max} \gtrsim \sqrt{D_{b-1}/D_b}$, then Theorem 4 suggests that QMC recovers the modified QAA's quadratic speedup. In such a scenario, however, it is likely that QAA itself

also favorably localizes on configurations which are close in Hamming distance to solutions of size $\geq b$. In such a situation, adding a large λ to the QAA likely does *not* enhance its performance, because it already benefits from (exponentially) favorable localization in the absence of λ. In other words, the only scenario where QMC can recover the QAA's quadratic speedup is one in which the quadratic speedup is irrelevant due to favorable localization, which can be exploited by both QAA and QMC. We note also that there is no reason *a priori* to expect such favorable localization to occur (and indeed, Fig. 13(a) of Appendix E 2 suggests that it typically does not), although we cannot strictly exclude it from formal arguments.

Proof. As before, we will use the Cheeger inequality to prove an upper bound on the spectral gap of the QMC Markov chain Δ_{QMC}. This gives us a lower bound on the QMC runtime via Eq. (A8). We will adopt identical notation and similar techniques to the parallel tempering proof in Appendix A 2. As in Eq. (A26), let $\mathcal{S}$ be the set of configurations with $H_{\mathrm{cost}}(z_i) > -\delta b$ for all i. Let ∂S_i represent the configurations $z_i \in S_i$ for which QMC can transition into S_i^c in a single update of at most k spins.

We will first consider the regime where $\pi_{\mathcal{S}} < 1/2$. We can compute

$$\begin{aligned} Q_{\mathcal{S},\mathcal{S}^c} &= \sum_{\substack{z_1\dots z_M\in\mathcal{S}\\ z_1'\dots z_M'\in\mathcal{S}^c}} \pi_{z_1\dots z_M} P_{z_1\dots z_M, z_1'\dots z_M'} \\ &\leq \sum_{i=1}^{M} \sum_{\substack{z_i\in\partial S_i\\ z_j\in S_j, j\neq i}} \pi_{z_1\dots z_M} \\ &\leq M \sum_{\substack{z_1\in\partial S_1\\ z_j\in S_j, j\neq 1}} \pi_{z_1\dots z_M}. \end{aligned} \tag{A50}$$

In the second line we used the fact that if $z_1\dots z_M \in \mathcal{S}$ can transition into $\mathcal{S}^c$, then $z_i \in \partial S_i$ for at least one replica i. The third line uses the standard cyclic permutation property of QMC Gibbs populations. We note that strictly speaking, one can choose to Trotterize the path integral in QMC in such a way that the cyclic permutation property is modified. For instance, if instead of the $H_{\mathrm{o.d.}}/H_{\mathrm{d}}$ decomposition above, we apply H_{cost}, H_ℓ and H_q in separate imaginary time slices, the QMC Gibbs weights will only be invariant under "even" cyclic shifts $z_i \to z_{i+2a \mod M}$ for $a \in \mathbb{Z}$. This does not affect the result because in such cases, the transition between $\mathcal{S}$ and $\mathcal{S}^c$ must still happen in one "block" of the cycle (e.g. one $H_{\mathrm{cost}}, H_\ell, H_q$ block in this example), and all the configurations within a single such block must be within a constant Hamming distance from each other. Finally, we remark that if at most $m < M$ Trotter slices are modified during a QMC update, then the factor of M in Eq. (A50) can be replaced with m. This can be seen by writing the $Q_{\mathcal{S},\mathcal{S}^c}$ as a sum over proposed updates to m replicas, then only summing over configurations with one of those m replicas i in ∂S_i.

Therefore, we have

$$\frac{Q_{\mathcal{S},\mathcal{S}^c}}{\pi_{\mathcal{S}}} \leq M \frac{\sum_{\substack{z_1\in\partial S_1\\ z_j\in S_j, j\neq 1}} \pi_{z_1\dots z_M}}{\sum_{z_1\dots z_M\in\mathcal{S}} \pi_{z_1\dots z_M}}. \tag{A51}$$

Now notice that the summations in the numerator and denominator of Eq. (A51) are only over configurations in $\mathcal{S}$. Thus, they can be related to the Gibbs state of $H_{\mathrm{QAA}} + \lambda H_\ell$ in a restricted Hilbert space that includes no configurations in $\mathcal{S}^c$. We will denote quantities in this restricted Hilbert space with a superscript (r,b), so that $H^{(r,b)} = H_{\mathrm{QAA}}^{(r,b)} + \lambda H_\ell^{(r,b)}$. Note now that because $H^{(r,b)}$ is identical to H on this restricted space, the populations that one would compute with QMC in this restricted space are related to their values in the full Hilbert space by an overall normalization factor:

$$\pi_{z_1\dots z_M}^{(r,b)} = \frac{\mathcal{Z}_\beta}{\mathcal{Z}_\beta^{(r,b)}} \pi_{z_1\dots z_M}. \tag{A52}$$

As a result, we may write

$$\begin{aligned} \frac{Q_{\mathcal{S},\mathcal{S}^c}}{\pi_{\mathcal{S}}} &\leq M \frac{\sum_{\substack{z_1\in\partial S_1\\ z_j\in S_j, j\neq 1}} \pi_{z_1\dots z_M}}{\sum_{z_1\dots z_M\in\mathcal{S}} \pi_{z_1\dots z_M}} \\ &= M \frac{\sum_{\substack{z_1\in\partial S_1\\ z_j\in S_j, j\neq 1}} \pi_{z_1\dots z_M}^{(r,b)}}{\sum_{z_1\dots z_M\in\mathcal{S}} \pi_{z_1\dots z_M}^{(r,b)}} \\ &= M \frac{\sum_{z_1\in\partial S_1} \pi_{z_1}^{(r,b)}}{\sum_{z_1\in S_1} \pi_{z_1}^{(r,b)}} \\ &\leq M k n^k \frac{e_{\max}^{(r,b)} D_b}{D_{b-1}}. \end{aligned} \tag{A53}$$

In the final line, we have used that there are $(k)^2\binom{n}{k} D_b \leq k' n^k D_b$ configurations in ∂S_1, and the definition $e_{\max}^{(r,b)} = \pi_{z_{\max}}^{(r,b)} D_{b-1}$. From Eq. (A53), we can thus immediately obtain the bound in Eq. (A49).

The discussion so far has assumed $\pi_{\mathcal{S}} < 1/2$. When $\pi_{\mathcal{S}} > 1/2$, we must instead compute $Q_{\mathcal{S},\mathcal{S}^c}/\pi_{\mathcal{S}^c}$ for the Cheeger bound in Eq. (A9). We multiply this quantity by π_α to obtain the quantity that appears in the QMC runtime definition in Eq. (A47),

$$\frac{Q_{\mathcal{S},\mathcal{S}^c}\pi_\alpha}{\pi_{\mathcal{S}^c}} \leq Q_{\mathcal{S},\mathcal{S}^c} \leq M \sum_{\substack{z_1\in\partial S_1\\ z_j\in S_j, j\neq 1}} \pi_{z_1\dots z_M}. \tag{A54}$$

By the above arguments, we may then write

$$Q_{\mathcal{S},\mathcal{S}^c} \leq M \frac{\mathcal{Z}_\beta^{(r,b)}}{\mathcal{Z}_\beta} \sum_{z_1\in\partial S_1} \pi_{z_1}^{(r,b)}. \tag{A55}$$

We now note that $\mathcal{Z}_\beta^{(r,b)} \leq \mathcal{Z}_\beta$, because the Gibbs weights contained in $\mathcal{Z}_\beta^{(r,b)}$ are a subset of the weights contained

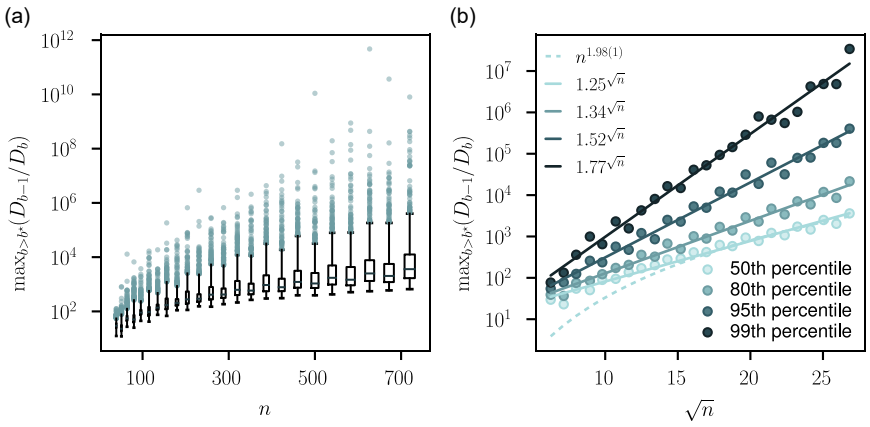

FIG. 7. Classical Markov chain runtime versus system size. (a) A box-and-whiskers plot of the classical runtime lower bound versus the system size n. The box endpoints are the 25th and 75th percentiles, and the whiskers are the 0th and 95th percentiles. (b) The runtime lower bounds at the 80th, 95th and 99th percentiles scale exponentially in $\sqrt{n}$. The 50th percentile runtime lower bound is also consistent with exponential scaling in $\sqrt{n}$.

in $\mathcal{Z}_\beta$. Note that we use the fact that the Hamiltonian does not have a sign problem, which ensures the positivity of the Gibbs weights. Thus, we have

$$\frac{Q_{\mathcal{S},\mathcal{S}^c}\pi_\alpha}{\pi_{\mathcal{S}^c}} \leq Mkn^k \frac{e_{\max}^{(r,b)} D_b}{D_{b-1}}, \tag{A56}$$

using the same reasoning as in Eq. (A53). As our bounds hold at any point during the adiabatic ramp and at any temperature $1/\beta$, we have thus shown Theorem 4.

Appendix B: Runtime scaling with system size

Here, we numerically study the runtime lower bounds for the classical Markov Chain Monte Carlo algorithms studied in Appendix A as a function of the number of vertices n, and compare the bounds against leading exact classical algorithms. The runtime lower bounds for these algorithms are equal to the quantity $\max_{b>b^\star}(D_{b-1}/D_b)$ up to polynomial factors in $1/n$, where D_b is the number of independent sets of size b, and $b^\star$ is the cutoff independent set size as defined in Appendix A 1. This quantity is large when there are many independent sets of some size $b-1$ compared to independent sets of size b. We are interested in determining how this quantity scales with n.

We randomly generate unit-disk graph instances with up to 720 vertices embedded on a two-dimensional square lattice with random 80% filling (see Fig. 1(a), main text). We study 1000 instances at each system size and compute $\max_{b>b^\star}(D_{b-1}/D_b)$ using the tensor-network algorithm for computing solution-space properties of combinatorial optimization problems detailed in Ref. [40]. We find that the independence polynomial of every single instance is *unimodal*, i.e., $D_0 \leq D_1 \leq \cdots \leq D_{b^\star} \geq \cdots \geq D_{\alpha-1} \geq D_\alpha$, which may be of independent interest [67]. This means that for the unit-disk graphs we study, in practice it is not strictly necessary to have a cutoff independent set size $b^\star$ in the runtime lower bound $\max_{b>b^\star}(D_{b-1}/D_b)$: any b with $D_{b-1}/D_b \geq 1$ can be used in the maximization. The vast majority (99.87%) of instances we study have $\max_b(D_{b-1}/D_b) = D_{\alpha-1}/D_\alpha$, and the remainder have $\max_b(D_{b-1}/D_b) = D_{\alpha-2}/D_{\alpha-1}$.

Figure 7(a) shows a box-and-whiskers plot of the full distribution of runtime lower bounds as a function of n. The variance of runtimes spans several orders of magnitude and increases with n, and the largest runtime over all the studied graphs is nearly 10^{12}. In Fig. 7(b), we plot various percentiles of $\max_{b>b^\star}(D_{b-1}/D_b)$ versus $\sqrt{n}$. We find that the runtime is exponential in $\sqrt{n}$ for instances in the 80th percentile and above. The 50th percentile runtime also appears to scale exponentially in $\sqrt{n}$ rather than polynomially. Therefore, the classical runtime lower bounds are (sub)exponentially faster than black-box search, which has an expected runtime of $\mathcal{O}(2^n/D_\alpha)$, which is exponential in n instead of $\sqrt{n}$.

We can compare the scaling of the runtime lower bound with system size to leading exact classical algorithms, which are guaranteed to return the largest independent set. The best exact classical algorithms for solving the unit-disk Maximum Independent Set problem find the solution in time $\mathcal{O}(c^{\sqrt{n}})$, for some constant $c \in (1,2)$. This scaling can be achieved using dynamic programming [68] or tensor-network methods [40]. Numerical evidence for the system sizes studied (see Fig. 2 in the main text) suggests that the actual SA runtime is linearly related to the SA runtime lower bound, suggesting that the typical SA runtime also scales as $\mathcal{O}(c^{\sqrt{n}})$. If this result holds as $n \to \infty$, then the scaling of both classical Markov chain algorithms and the modified QAA are typically polynomially related to the best classical algorithms. In particular, if the SA runtime scaling is $\mathcal{O}(c^{\sqrt{n}})$, then the runtime of our modified QAA scales roughly as $\mathcal{O}(\sqrt{c}^{\sqrt{n}})$.

Appendix C: Resolvent method for the minimum gap

1. Derivation of the minimum gap formula

Here we will derive an exact method to pertubatively compute the minimum gap Δ_{QAA} of $H_{\mathrm{QAA}} = H_{\mathrm{cost}} - H_q$ when the avoided level crossing location $(\Omega/\delta)_\star \ll 1$. In the main text we used degenerate perturbation theory to compute, to leading order in Ω/δ, the orthogonal states $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ which approximate the ground and first excited eigenstates at $\Omega/\delta \lesssim (\Omega/\delta)_\star \ll 1$ (see Eq. (11), main text). Here we will exactly compute Δ_{QAA} in terms of the matrix elements of an effective Hamiltonian $H_{\mathrm{eff}}(z)$ acting on the subspace spanned by $|\mathcal{G}\rangle, |\mathcal{E}\rangle$, defined by the projector $P = |\mathcal{G}\rangle\langle\mathcal{G}| + |\mathcal{E}\rangle\langle\mathcal{E}|$. Our main results are in Eq. (C6), which gives Δ_{QAA} exactly in terms of the matrix elements of $H_{\mathrm{eff}}(z)$, and Eq. (C12), which simplifies the result under a motivated approximation.

$H_{\mathrm{eff}}(z)$ can be derived by rewriting the eigenvalue equation $H_{\mathrm{QAA}}|\psi\rangle = z|\psi\rangle$ as $H_{\mathrm{QAA}}(P+Q)|\psi\rangle = z|\psi\rangle$, where $Q = \mathbb{1} - P$, then multiplying by P and Q to obtain a system of equations for the eigenvector $|\psi\rangle$:

$$\begin{bmatrix} QH_{\mathrm{QAA}}Q & QH_{\mathrm{QAA}}P \\ PH_{\mathrm{QAA}}Q & PH_{\mathrm{QAA}}P \end{bmatrix} \begin{bmatrix} Q|\psi\rangle \\ P|\psi\rangle \end{bmatrix} = z \begin{bmatrix} Q|\psi\rangle \\ P|\psi\rangle \end{bmatrix}. \tag{C1}$$

These equations can then be written in terms of $P|\psi\rangle$ as

$$\underbrace{\left[PH_{\mathrm{QAA}}P + PH_{\mathrm{QAA}}\frac{Q}{z - QH_{\mathrm{QAA}}Q}H_{\mathrm{QAA}}P\right]}_{H_{\mathrm{eff}}(z)}|\psi\rangle = zP|\psi\rangle. \tag{C2}$$

The left hand side of the equation defines $H_{\mathrm{eff}}(z)$, the effective Hamiltonian in the subspace spanned by $|\mathcal{G}\rangle, |\mathcal{E}\rangle$. The second term in $H_{\mathrm{eff}}(z)$ can be interpreted as a perturbative addition to original Hamiltonian, $PH_{\mathrm{QAA}}P$, due to higher-order couplings in Ω/δ that come from the Q subspace, which is energetically separated from the P subspace. Expanding the denominator using the matrix Taylor expansion $(A+B)^{-1} = A^{-1}\sum_{l=0}^{\infty}(-BA^{-1})^l$, we receive

$$H_{\mathrm{eff}}(z) = PH_{\mathrm{cost}}P - \sum_{l=0}^{\infty} P\left(-H_q\frac{Q}{z - H_{\mathrm{cost}}}\right)^l H_q P, \tag{C3}$$

where we have used $PH_{\mathrm{cost}}Q = 0$ because $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ are eigenstates of H_{cost}. This form of $H_{\mathrm{eff}}(z)$ has an intuitive interpretation: each order l applies a factor of H_q, but is suppressed by a factor of $\mathcal{O}(\Omega/\delta)$.

Prior works have estimated Δ_{QAA} from the off-diagonal matrix element of $H_{\mathrm{eff}}(E_\star)$ evaluated at $(\Omega/\delta)_\star$ as [25, 44]

$$\tilde{\Delta}_{\mathrm{QAA}} = 2|\langle\mathcal{G}|H_{\mathrm{eff}}(E_\star)|\mathcal{E}\rangle|, \tag{C4}$$

which we analyzed in the main text [see Eq. (8)]. This equation has an intuitive interpretation under the assumption of Landau-Zener physics on $H_{\mathrm{eff}}(z)$, which we illustrate in Fig. 8(a). At $\Omega/\delta = 0$, $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ are eigenstates of H_{QAA} with eigenenergies given by the on-diagonal entries of $H_{\mathrm{eff}}(z)$ ($\langle\mathcal{G}|H_{\mathrm{cost}}|\mathcal{G}\rangle$ and $\langle\mathcal{E}|H_{\mathrm{cost}}|\mathcal{E}\rangle$, respectively). At the avoided level crossing $\Omega/\delta = (\Omega/\delta)_\star$, we expect the on-diagonal eigenenergies of $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ in $H_{\mathrm{eff}}(z)$ to cross at a value close to $E_\star$ for some value of $z \simeq E_\star$, which we denote by z'. The gap of $H_{\mathrm{eff}}(z')$ at $(\Omega/\delta)_\star$ is then given by the off-diagonal coupling $2|\langle\mathcal{G}|H_{\mathrm{eff}}(z')|\mathcal{E}\rangle| \simeq \tilde{\Delta}_{\mathrm{QAA}}$. $\tilde{\Delta}_{\mathrm{QAA}}$ indeed captures the correct qualitative physics, but is quantitatively inaccurate. Here we show that Δ_{QAA} can be computed exactly from the matrix elements of $H_{\mathrm{eff}}(z)$ in Eqs. (C6) and (C12).

$\tilde{\Delta}_{\mathrm{QAA}}$ does not equal Δ_{QAA} in general because of the z-dependence of $H_{\mathrm{eff}}(z)$, which prevents it from being interpreted as a true Hamiltonian. The only guaranteed relationship between H_{eff} and the spectrum of H_{QAA} is that each eigenvalue z of H_{QAA} is also an eigenvalue of $H_{\mathrm{eff}}(z)$ [see Eq. (C2)], i.e.,

$$\det[z - H_{\mathrm{eff}}(z)] = 0 \tag{C5}$$

whenever z is an eigenvalue of H_{QAA}. Δ_{QAA} can therefore be obtained exactly from taking the difference between the first two values of z that solve Eq. (C5), which are the two lowest energy eigenvalues at $z = E_\star, E_\star + \Delta_{\mathrm{QAA}}$. We show an example of numerically using this method to exactly reconstruct Δ_{QAA} in Fig. 8(b) for a star graph with $b = 40$ branches of length $\ell = 2$. In contrast, we find that $\tilde{\Delta}_{\mathrm{QAA}}$, computed numerically, overestimates Δ_{QAA} for the same instance by a factor of 4.53 (Fig 8(b), inset). This discrepancy is due to the z-dependence of H_{eff}, which we show in Fig. 8(c) for the same instance.

To account for this z-dependence, we will consider z in the neighborhood of $E_\star$, and compute the leading order, linear dependence of H_{eff} on z. We adopt the following ansatz by expanding $H_{\mathrm{eff}}(z)$ around a reference point $z = z_0$:

$$H_{\mathrm{eff}}(z) = \begin{bmatrix} \langle\mathcal{E}|H_{\mathrm{eff}}(z_0)|\mathcal{E}\rangle + m_{ee}(z - z_0) & \langle\mathcal{E}|H_{\mathrm{eff}}(z_0)|\mathcal{G}\rangle + m_{ge}(z - z_0) \\ \langle\mathcal{E}|H_{\mathrm{eff}}(z_0)|\mathcal{G}\rangle + m_{ge}(z - z_0) & \langle\mathcal{G}|H_{\mathrm{eff}}(z_0)|\mathcal{G}\rangle + m_{gg}(z - z_0) \end{bmatrix}, \tag{C6}$$

where $m_{ge} = m_{eg}$ because H_{eff} is real. Δ_{QAA} can then be obtained from solving Eq. (C5) using the ansatz for

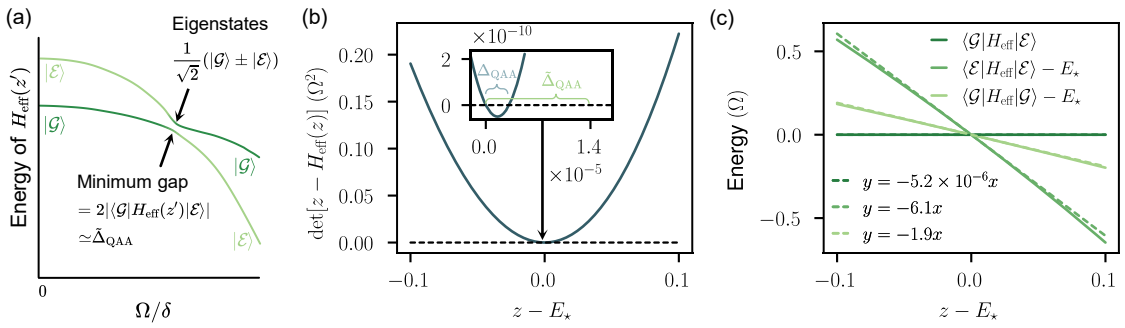

FIG. 8. Computing the minimum gap using the resolvent formalism. (a) When the avoided level crossing location $(\Omega/\delta)_\star \ll 1$, the avoided level crossing can be understood in terms of Landau-Zener physics between $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ under $H_{\rm eff}(z')$. At $\Omega/\delta = (\Omega/\delta)_\star$, $|\mathcal{E}\rangle$ and $|\mathcal{G}\rangle$ have the same on-diagonal energy under $H_{\rm eff}(z)$, and the minimum gap of $H_{\rm eff}(z')$ is given by their off-diagonal coupling $2|\langle\mathcal{G}|H_{\rm eff}(z')|\mathcal{E}\rangle|$. (b) $\Delta_{\rm QAA}$ equals the difference of the first two zeroes of $\det[z - H_{\rm eff}(z)]$, which occur at $z = E_\star$ and $E_\star + \Delta_{\rm QAA}$ (light blue, inset). The estimated gap $\tilde{\Delta}_{\rm QAA} = 2|\langle\mathcal{G}|H_{\rm eff}(E_\star)|\mathcal{E}\rangle|$ overestimates the minimum gap $\Delta_{\rm QAA}$ by a factor of 4.53 for this instance (light green, inset). (c) When $z - E_\star$ is small, matrix elements of $H_{\rm eff}$ (solid lines) are well-approximated by a linear function of z (dashed lines). For the star graph with $b = 40, \ell = 2$, the $\langle\mathcal{G}|H_{\rm eff}(z)|\mathcal{E}\rangle$ matrix element changes as a function of $z - E_\star$ with a slope of $m_{ge} = -5.2 \times 10^{-6}$. The matrix elements $\langle\mathcal{G}|H_{\rm eff}(z)|\mathcal{G}\rangle$ and $\langle\mathcal{E}|H_{\rm eff}(z)|\mathcal{E}\rangle$ change at much higher rates of $m_{gg} = -1.9$ and $m_{ee} = -6.1$, respectively.

$H_{\rm eff}$ in Eq. (C6), which gives

$$\Delta_{\rm QAA} = 2\Big[\langle\mathcal{E}|H_{\rm eff}(z_0)|\mathcal{G}\rangle^2 f_{ee}f_{gg} + \frac{1}{4}\left[(f_{ee}+f_{gg})\Delta\bar{E} + (f_{gg}-f_{ee})(\bar{E}-z_0)\right]^2 + \langle\mathcal{E}|H_{\rm eff}(z_0)|\mathcal{G}\rangle m_{ge}\left[(f_{ee}+f_{gg})(\bar{E}-z_0) + (f_{gg}-f_{ee})\Delta\bar{E}\right] + m_{ge}^2[(\bar{E}-z_0)^2 - \Delta\bar{E}^2]\Big]^{1/2} / \left[f_{gg}f_{ee} - m_{ge}^2\right], \tag{C7}$$

where we have defined the mean and difference of the on-diagonal energies,

$$\bar{E}(z_0) = \frac{1}{2}(\langle\mathcal{E}|H_{\rm eff}(z_0)|\mathcal{E}\rangle + \langle\mathcal{G}|H_{\rm eff}(z_0)|\mathcal{G}\rangle)$$
$$\Delta\bar{E}(z_0) = \frac{1}{2}(\langle\mathcal{E}|H_{\rm eff}(z_0)|\mathcal{E}\rangle - \langle\mathcal{G}|H_{\rm eff}(z_0)|\mathcal{G}\rangle), \tag{C8}$$

and let

$$f_{ee} = 1 - m_{ee}$$
$$f_{gg} = 1 - m_{gg}. \tag{C9}$$

Eq. (C7) therefore gives $\Delta_{\rm QAA}$ in terms of the matrix elements of $H_{\rm eff}$ and their first-order derivatives in z. In the absence of z-dependence ($m_{ee} = m_{gg} = m_{ge} = 0$), one can check that this expression reduces the result one would obtain from directly diagonalizing $H_{\rm eff}(z_0)$. Therefore, as expected, when $H_{\rm eff}$ is independent of z, it can be treated as a true Hamiltonian acting on $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ and diagonalized to find $\Delta_{\rm QAA}$.

Although Eq. (C7) is exact, we can vastly simplify it using intuition from Landau-Zener theory. Suppose, to good approximation, there exists a z' such that the diagonal entries of $H_{\rm eff}(z')$ intersect at z' for $\Omega/\delta = (\Omega/\delta)_\star$: $\langle\mathcal{E}|H_{\rm eff}(z')|\mathcal{E}\rangle = \langle\mathcal{G}|H_{\rm eff}(z')|\mathcal{G}\rangle = z'$. Because $H_{\rm eff}$ is independent of the point of expansion z_0 in the regime where the linear approximation is valid, we may choose $z_0 = z'$. Using our assumption that the diagonal entries of $H_{\rm eff}(z')$ intersect at $(\Omega/\delta)_\star$, we then have $\Delta\bar{E}(z_0) = 0$ and $\bar{E}(z_0) = z_0$. Under this choice of z_0, $\Delta_{\rm QAA}$ simplifies to

$$\Delta_{\rm QAA} = \frac{2\sqrt{\langle\mathcal{E}|H_{\rm eff}(z')|\mathcal{G}\rangle^2 f_{ee}f_{gg}}}{f_{gg}f_{ee} - m_{ge}^2}. \tag{C10}$$

We may further simplify this expression using the fact that we expect $|m_{ge}| \ll |f_{gg}|, |f_{ee}|$. To see this, we compute $dH_{\rm eff}/dz$ for $z \in \mathbb{R}$ as

$$\frac{dH_{\rm eff}}{dz} = -\Omega^2 P H_q \left(\frac{Q}{z - QH_{\rm QAA}Q}\right)^2 H_q P \tag{C11}$$
$$= -\Omega^2 \left[\frac{Q}{z - QH_{\rm QAA}Q}H_q P\right]^\dagger \left[\frac{Q}{z - QH_{\rm QAA}Q}H_q P\right],$$

which is similar to the second term of $H_{\text{eff}}(z)$ in Eq. (C2). By expanding Eq. (C11) in powers of Ω/δ, as in Eq. (C3), one can see that the on-diagonal entries can in general be large because they connect either $|\mathcal{G}\rangle$ or $|\mathcal{E}\rangle$ to itself via even multiples of H_q. On the other hand, the off-diagonal entries should be smaller by $\mathcal{O}(\Delta_{\text{QAA}})$ because they connect $|\mathcal{G}\rangle$ to $|\mathcal{E}\rangle$ via odd multiples of H_q, similar to the off-diagonal entries of $H_{\text{eff}}(z)$. Therefore, we expect that $|m_{ge}|/|f_{ee}|, |m_{ge}|/|f_{gg}| = \mathcal{O}(\Delta_{\text{QAA}})$. We verify numerically that $|m_{ge}| = \mathcal{O}(\Delta_{\text{QAA}})$ and $f_{gg}, f_{ee} = \mathcal{O}(1)$ in Fig. 8(c) for an example star graph. Therefore, to good approximation we have

$$\Delta_{\text{QAA}} = \frac{2|\langle\mathcal{E}| H_{\text{eff}}(E_\star) |\mathcal{G}\rangle|}{\sqrt{f_{gg} f_{ee}}} = \frac{2\tilde{\Delta}_{\text{QAA}}}{\sqrt{f_{gg} f_{ee}}}. \tag{C12}$$

Note that by the form of Eq. (C11), dH_{eff}/dz can be written as $-\Omega^2$ times a positive semidefinite operator, so all the derivatives of H_{eff} are negative. Therefore, $f_{ee}, f_{gg} \geq 1$, so $\tilde{\Delta}_{\text{QAA}}$ is an overestimate of the gap, consistent with our numerical results on the star graph in Fig. 8(b). We verify numerically in Appendix D 1, Fig. 10(d) that Eq. (C12) recovers the Δ_{QAA} for the star graph to high accuracy.

2. Validity of the resolvent method

For $\tilde{\Delta}_{\text{QAA}}$ to be a good qualitative predictor of Δ_{QAA} via Eq. (C12), the factors f_{gg}, f_{ee} cannot be large compared to the minimum gap as to change its leading-order scaling behavior with n. The z-dependence of $H_{\text{eff}}(z)$ comes from the factor of $(z - QH_{\text{QAA}}Q)^{-1}$ in Eq. (C2), which creates a pole at every eigenvalue of $QH_{\text{QAA}}Q$. Although this creates significant z-dependence in $H_{\text{eff}}(E_\star)$ if $QH_{\text{QAA}}Q$ has an eigenvalue close to the ground state energy $E_\star \equiv E_0$, the z-dependence will be modest if the ground state energy of $QH_{\text{QAA}}Q$ is significantly larger than E_0. We expect this to occur when Q is a sufficiently good projector out of the ground and first-excited states of H_{QAA}, and the second-excited state energy of H_{QAA} is much larger than Δ_{QAA}. To formalize this intuition, in the following Theorem 5 we relate the energy difference between E_0 and the ground state energy of $QH_{\text{QAA}}Q$, denoted δE, to the overlap of $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ with the ground and first excited eigenstates of H_{QAA}.

Theorem 5. *Denote the eigenstates of H_{QAA} at $(\Omega/\delta)_\star$ as $|\psi_0\rangle, |\psi_1\rangle \ldots, |\psi_{2^n}\rangle$ with corresponding eigenvalues $E_0 \leq E_1 \leq \cdots \leq E_{2^n}$, and assume that $\Delta_{\text{QAA}} = E_1 - E_0 \ll E_2 - E_1$. Denote the ground state energy of $QH_{\text{QAA}}Q$ by $E_0 + \delta E$. Then, if $|\langle\psi_0|\mathcal{G}\rangle\langle\psi_1|\mathcal{E}\rangle - \langle\psi_0|\mathcal{E}\rangle\langle\psi_1|\mathcal{G}\rangle|^2 \gg \Delta_{\text{QAA}}$, δE is bounded as*

$$\delta E \geq \frac{1}{4}(E_2 - E_1)|\langle\psi_0|\mathcal{G}\rangle\langle\psi_1|\mathcal{E}\rangle - \langle\psi_0|\mathcal{E}\rangle\langle\psi_1|\mathcal{G}\rangle|^2. \tag{C13}$$

Hence, if $|\langle\psi_0|\mathcal{G}\rangle\langle\psi_1|\mathcal{E}\rangle - \langle\psi_0|\mathcal{E}\rangle\langle\psi_1|\mathcal{G}\rangle|$ and $E_2 - E_1$ are at worst polynomially small in n, we will have δE at worst polynomially small in n. This will make the correction factors f_{ee}, f_{gg} at most polynomially large in n, and thus subleading when Δ_{QAA} is exponentially small in n. The quantity $\langle\psi_0|\mathcal{G}\rangle\langle\psi_1|\mathcal{E}\rangle - \langle\psi_0|\mathcal{E}\rangle\langle\psi_1|\mathcal{G}\rangle$ can be interpreted as the area of the parallelogram defined by $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ in the $|\psi_0\rangle, |\psi_1\rangle$ subspace. If we define $\mathcal{P} = |\psi_0\rangle\langle\psi_0| + |\psi_1\rangle\langle\psi_1|$, the condition that $(\langle\psi_0|\mathcal{G}\rangle\langle\psi_1|\mathcal{E}\rangle - \langle\psi_0|\mathcal{E}\rangle\langle\psi_1|\mathcal{G}\rangle)$ is large is thus both a statement about the size of the overlaps $\langle\mathcal{G}|\mathcal{P}|\mathcal{G}\rangle, \langle\mathcal{E}|\mathcal{P}|\mathcal{E}\rangle$ and also a statement about the linear independence of $\mathcal{P}|\mathcal{G}\rangle, \mathcal{P}|\mathcal{E}\rangle$. Intuitively, $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ must have good overlap with the span of $|\psi_0\rangle$ and $|\psi_1\rangle$, and must furthermore capture sufficiently different directions within this space. We expect $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ to satisfy this condition when $(\Omega/\delta)_\star \ll 1$ because $|\mathcal{G}\rangle$ approximates $|\psi_0\rangle$ and $|\mathcal{E}\rangle$ approximates $|\psi_1\rangle$ in perturbation theory, and $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ are orthogonal.

Proof. We first find the ground state energy of $QH_{\text{QAA}}Q$ by computing $\min_\phi \langle\phi| H_{\text{QAA}} |\phi\rangle$ subject to $\langle\phi|\phi\rangle = 1$ and $P|\phi\rangle = 0$. This can be formulated as the minimization of $\langle\phi| H_{\text{QAA}} |\phi\rangle - \zeta_0\langle\phi|\mathcal{G}\rangle - \zeta_1\langle\phi|\mathcal{E}\rangle - \mu(\langle\phi|\phi\rangle - 1)$ where $\zeta_{0,1}, \mu$ are Lagrange multipliers. Writing $|\phi\rangle$ in the eigenbasis $|\psi_i\rangle$ of H_{QAA} and setting the derivatives with respect to $\langle\phi|\psi_i\rangle$ of this expression to zero yields the condition

$$\langle\psi_i|\phi\rangle = \frac{1}{2}\frac{1}{E_i - \mu}(\zeta_0\langle\psi_i|\mathcal{G}\rangle + \zeta_1\langle\psi_i|\mathcal{E}\rangle). \tag{C14}$$

Plugging Eq. (C14) into the constraints $\langle\mathcal{G}|\phi\rangle = \langle\mathcal{E}|\phi\rangle = 0$ yields two equations involving $\zeta_{0,1}, \mu$, and $\langle\psi_i|\mathcal{G}\rangle, \langle\psi_i|\mathcal{E}\rangle$. Solving one equation for ζ_0 and substituting into the other yields, after simplification,

$$\begin{aligned} 0 &= \sum_{i,j} \frac{|\langle\psi_i|\mathcal{G}\rangle\langle\psi_j|\mathcal{E}\rangle - \langle\psi_i|\mathcal{E}\rangle\langle\psi_j|\mathcal{G}\rangle|^2}{(E_i - \mu)(E_j - \mu)} \\ &= \sum_{ij} \frac{|\langle\psi_i\psi_j|\Phi\rangle|^2}{(E_i - \mu)(E_j - \mu)}, \end{aligned} \tag{C15}$$

where we have defined the wavefunction $|\Phi\rangle = |\mathcal{G}\rangle\otimes|\mathcal{E}\rangle - |\mathcal{E}\rangle\otimes|\mathcal{G}\rangle$, which exists in a doubled Hilbert space. Now plugging Eq. (C14) into $\langle\phi| H_{\text{QAA}} |\phi\rangle / \langle\phi|\phi\rangle$ and simplifying using Eq. (C15) yields the conclusion that the minimum energy of $QH_{\text{QAA}}Q$ is μ. So we must use Eq. (C15), which gives a constraint on μ, to constrain $\delta E = \mu - E_0$.

Before doing so we briefly note that although we must have $\mu \geq E_0$ by definition, we must exclude the possibility that $E_0 \leq \mu \leq E_1$ by contradiction: were this to happen, we could rewrite Eq. (C15) as

$$\begin{aligned} &2\frac{|\langle\psi_0\psi_1|\Phi\rangle|^2}{(\mu - E_0)(E_1 - \mu)} + 2\frac{1}{\mu - E_0}\sum_{i\neq 0,1}\frac{|\langle\psi_0\psi_i|\Phi\rangle|^2}{E_i - \mu} \\ &= 2\frac{1}{E_1 - \mu}\sum_{i\neq 0,1}\frac{|\langle\psi_1\psi_i|\Phi\rangle|^2}{E_i - \mu} + \sum_{i,j\neq 0,1}\frac{|\langle\psi_i\psi_j|\Phi\rangle|^2}{(E_i - \mu)(E_j - \mu)}. \end{aligned} \tag{C16}$$

Here all terms are positive, but the first term on the left hand side is $\mathcal{O}(|\langle\psi_0\psi_1|\Phi\rangle|^2\Delta_{\mathrm{QAA}}^{-2})$ (because $\mu - E_0, E_1 \leq E_1 - E_0 = \Delta_{\mathrm{QAA}}$), whereas the first term on the right hand side is only $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-1})$. The equation is thus impossible to satisfy under the assumption $|\langle\psi_0|\mathcal{G}\rangle\langle\psi_1|\mathcal{E}\rangle - \langle\psi_1|\mathcal{G}\rangle\langle\psi_0|\mathcal{E}\rangle|^2 \gg \Delta_{\mathrm{QAA}}$.

Therefore, to constrain δE, we can assume that $E_1 \leq \mu \leq E_2$, and use Eq. (C15) to write

$$\begin{aligned}\frac{|\langle\psi_0\psi_1|\Phi\rangle|^2 + |\langle\psi_1\psi_0|\Phi\rangle|^2}{(\mu-E_0)(\mu-E_1)} &\leq \frac{1}{\mu-E_0}\sum_{i\neq 0,1}\frac{|\langle\psi_0\psi_i|\Phi\rangle|^2 + |\langle\psi_i\psi_0|\Phi\rangle|^2}{E_i-\mu} + \frac{1}{\mu-E_1}\sum_{i\neq 0,1}\frac{|\langle\psi_1\psi_i|\Phi\rangle|^2 + |\langle\psi_i\psi_1|\Phi\rangle|^2}{E_i-\mu} \\ &\leq \frac{1}{(\mu-E_0)(E_2-\mu)}(2|\langle\psi_0|\mathcal{G}\rangle|^2 + 2|\langle\psi_0|\mathcal{E}\rangle|^2 - |\langle\psi_0\psi_1|\Phi\rangle|^2 - |\langle\psi_1\psi_0|\Phi\rangle|^2) \\ &+ \frac{1}{(\mu-E_1)(E_2-\mu)}(2|\langle\psi_1|\mathcal{G}\rangle|^2 + 2|\langle\psi_1|\mathcal{E}\rangle|^2 - |\langle\psi_0\psi_1|\Phi\rangle|^2 - |\langle\psi_1\psi_0|\Phi\rangle|^2) \end{aligned} \tag{C17}$$

Between the first and second lines we used $E_2 - \mu \leq E_i - \mu$ for $i \geq 2$, and

$$\sum_{i\neq 0,1}|\langle\psi_0\psi_i|\Phi\rangle|^2 = |\langle\psi_0|\mathcal{G}\rangle|^2 + |\langle\psi_0|\mathcal{E}\rangle|^2 - |\langle\psi_0\psi_1|\Phi\rangle|^2. \tag{C18}$$

Now multiplying out all the denominators and using $E_0 = E_\star, E_1 = E_\star + \Delta_{\mathrm{QAA}}$, we obtain

$$\begin{aligned}\mu - E_\star \geq (E_2 - E_\star)&\frac{|\langle\psi_0\psi_1|\Phi\rangle|^2 + |\langle\psi_1\psi_0|\Phi\rangle|^2}{2|\langle\psi_0|\mathcal{G}\rangle|^2 + 2|\langle\psi_0|\mathcal{E}\rangle|^2 + 2|\langle\psi_1|\mathcal{G}\rangle|^2 + 2|\langle\psi_1|\mathcal{E}\rangle|^2 - (|\langle\psi_0\psi_1|\Phi\rangle|^2 + |\langle\psi_1\psi_0|\Phi\rangle|^2)} \\ + \Delta_{\mathrm{QAA}}&\frac{2|\langle\psi_0|\mathcal{G}\rangle|^2 + 2|\langle\psi_0|\mathcal{E}\rangle|^2 - (|\langle\psi_0\psi_1|\Phi\rangle|^2 + |\langle\psi_1\psi_0|\Phi\rangle|^2)}{2|\langle\psi_0|\mathcal{G}\rangle|^2 + 2|\langle\psi_0|\mathcal{E}\rangle|^2 + 2|\langle\psi_1|\mathcal{G}\rangle|^2 + 2|\langle\psi_1|\mathcal{E}\rangle|^2 - (|\langle\psi_0\psi_1|\Phi\rangle|^2 + |\langle\psi_1\psi_0|\Phi\rangle|^2)}. \end{aligned} \tag{C19}$$

The final term can be dropped because it is small, by the assumption $|\langle\psi_0|\mathcal{G}\rangle\langle\psi_1|\mathcal{E}\rangle - \langle\psi_1|\mathcal{G}\rangle\langle\psi_0|\mathcal{E}\rangle|^2 \gg \Delta_{\mathrm{QAA}}$. The bound in Theorem 5 then follows from maximizing the denominator in the first term.

3. Conditions for a perturbative avoided level crossing

By the arguments in Appendix C 2, the formula in Eq. (C7) for the minimum gap of $H_{\mathrm{QAA}} = H_{\mathrm{cost}} - H_q$ converges when the location of the avoided level crossing $(\Omega/\delta)_\star \ll 1$. Here we establish a condition for when this occurs, given in Eq. (C26), and motivate why we expect this condition to hold for problem instances with flat energy landscapes. We refer the reader to Ref. [69] for a detailed framework to predict $(\Omega/\delta)_\star$ for general combinatorial optimization problems.

Recall that the perturbed eigenstates (energy shifts) are the eigenvectors (eigenvalues) of the perturbed Hamiltonian

$$H^{(2)} = -\frac{\Omega^2}{\delta}\left(H_{se} + \sum_{u\in V} n_u - H_{fv}\right), \tag{C20}$$

where

$$H_{fv} = \sum_{u\in V}(\mathbb{1} - n_u)\prod_{(u,v)\in E}(\mathbb{1} - n_v), \tag{C21}$$

counts the number of free vertices for each independent set $|z\rangle$ (vertices which can be added to $|z\rangle$ without violating the independent set constraint). $|\mathcal{G}\rangle$ is the ground state of $H^{(2)}$ in the $H_{\mathrm{cost}} = -\delta\alpha$ manifold because this is the instantaneous ground state of the system as $\Omega/\delta \to 0$ (see Fig. 3(b), main text). Its perturbed energy under $H^{(2)}$ is

$$\langle\mathcal{G}|H_{\mathrm{cost}} + H^{(2)}|\mathcal{G}\rangle = -\delta\alpha - \frac{\Omega^2}{\delta}(\alpha + \langle\mathcal{G}|H_{se}|\mathcal{G}\rangle), \tag{C22}$$

where we have used that $\langle\mathcal{G}|H_{fv}|\mathcal{G}\rangle = 0$ because no vertices can be added to $|\mathcal{G}\rangle$. The last term counts the expected number of spin exchanges possible between neighboring vertices in $|\mathcal{G}\rangle$.

$|\mathcal{E}\rangle$ can be found by determining the H_{cost} manifold whose ground state energy intersects $|\mathcal{G}\rangle$ first at finite Ω/δ (see Fig. 3(b) and Ref. [44] for a discussion). Suppose $|\mathcal{E}\rangle$ is the ground state of Eq. (C20) in the $H_{\mathrm{cost}} = -\delta b$

manifold, for some unknown b. Then the perturbed eigenenergy of $|\mathcal{E}\rangle$ is

$$\langle\mathcal{E}|\, H_{\text{cost}} + H^{(2)}\,|\mathcal{E}\rangle = -\delta b - \frac{\Omega^2}{\delta}(b + \langle\mathcal{E}|\, H_{se}\,|\mathcal{E}\rangle - \langle\mathcal{E}|\, H_{fv}\,|\mathcal{E}\rangle). \quad \text{(C23)}$$

Note we have assumed that the ground state of each manifold of $H^{(2)}$ is nondegenerate, so that $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ can be uniquely identified. On instances where the degeneracy is not broken, or the energy splitting between the ground and first excited state of a manifold is too small to accurately identify $|\mathcal{G}\rangle$ or $|\mathcal{E}\rangle$, one can compute $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$ by going to higher order in perturbation theory.

$(\Omega/\delta)_\star$ can then be estimated by computing the value of Ω/δ where Eqs. (C22) and (C23) intersect, which is

$$(\Omega/\delta)_\star = \sqrt{\frac{\alpha - b}{\langle\mathcal{E}|H_{se}|\mathcal{E}\rangle - \langle\mathcal{G}|H_{se}|\mathcal{G}\rangle - \langle\mathcal{E}|H_{fv}|\mathcal{E}\rangle - \alpha + b}} \quad \text{(C24)}$$

to second order in Ω/δ. If $(\Omega/\delta)_\star \ll 1$, then the numerator of Eq. (C24) must be much smaller than the denominator:

$$\langle\mathcal{E}|\, H_{se}\,|\mathcal{E}\rangle - \langle\mathcal{G}|\, H_{se}\,|\mathcal{G}\rangle \gg 2(\alpha - b) + \langle\mathcal{E}|\, H_{fv}\,|\mathcal{E}\rangle\,. \quad \text{(C25)}$$

We can use the bound $\alpha - b \geq \langle\mathcal{E}|\, H_{fv}\,|\mathcal{E}\rangle$ to get the following condition for when $(\Omega/\delta)_\star \ll 1$:

$$\langle\mathcal{E}|\, H_{se}\,|\mathcal{E}\rangle - \langle\mathcal{G}|\, H_{se}\,|\mathcal{G}\rangle \gg 3(\alpha - b). \quad \text{(C26)}$$

Therefore, $(\Omega/\delta)_\star \ll 1$ when $|\mathcal{E}\rangle$ has a large number of expected possible spin exchanges compared to $|\mathcal{G}\rangle$, and $|\mathcal{E}\rangle$ is comprised of near-optimal independent sets. This is exactly the case on problem instances with flat energy landscapes at near-optimal independent set size $b \simeq \alpha$. Because there are many independent sets of size b with freedom to spin-exchange (see e.g. the configuration graph in Fig. 1, main text), we might expect $\langle\mathcal{E}|\, H_{se}\,|\mathcal{E}\rangle$ to be large. For example, if each vertex in independent sets of size b has k possible spin exchanges, then $\langle\mathcal{E}|\, H_{se}\,|\mathcal{E}\rangle = kb$ is extensively large with n. We similarly expect $\langle\mathcal{G}|\, H_{se}\,|\mathcal{G}\rangle = k'\alpha$, if vertices in independent sets of size α have k' possible spin exchanges (in the case where there is a unique largest independent set, $k' = 0$). Since there are far fewer independent sets of size α, and larger independent sets may have less freedom to spin-exchange under the independent set constraint, we might expect that $k > k'$ and therefore that $\langle\mathcal{E}|\, H_{se}\,|\mathcal{E}\rangle \gg \langle\mathcal{G}|\, H_{se}\,|\mathcal{G}\rangle$ for large systems. Therefore, on problem instances with flat energy landscapes, we expect the avoided level crossing location to occur near the end of the ramp, $(\Omega/\delta)_\star \ll 1$.

We verify that this interpretation is correct for the family of star graphs in Appendix D. We consider the case of fixed branch length ℓ, and look at the avoided level crossing location as the number of branches (and therefore n) grows. The largest independent set is unique, so $\langle\mathcal{G}|\, H_{se}\,|\mathcal{G}\rangle = 0$. $|\mathcal{E}\rangle$ is in the $H_{\text{cost}} = -\delta(\alpha - 1)$ manifold, so the right hand side of Eq. (C26) is equal to 3. Typical independent sets in $|\mathcal{E}\rangle$ can participate in $\mathcal{O}(n)$ spin exchanges. Therefore, by Eq. (C24) the avoided level crossing location $(\Omega/\delta)_\star$ goes like $\mathcal{O}(1/\sqrt{n_b}) = \mathcal{O}(1/\sqrt{n})$ as $n \to \infty$.

4. Experimental Rydberg Hamiltonian resolvent gaps

Here we analyze the performance of the Rydberg atom array experiment [23] using the resolvent gap formalism described in Appendix C 1. Because the Rydberg Hamiltonian H_{Ryd} (Eq. (3), main text) has long-range interactions not present in the Maximum Independent Set cost function H_{cost}, we must modify our perturbative formalism developed to predict the minimum gap Δ_{QAA}. Here we describe our method to perturbatively compute Δ_{QAA} for the Rydberg Hamiltonian. We then verify that the resolvent gap formalism qualitatively captures the experimental performance.

In the main text, we estimated the parameters of the avoided level crossing $|\mathcal{G}\rangle, |\mathcal{E}\rangle, (\Omega/\delta)_\star$, and $E_\star$ (see Fig. 3(b), main text) by solving for the perturbative Hamiltonian $H^{(2)}$ approximating the system Hamiltonian H_{QAA} at small Ω/δ (Eq. (11), main text). To find $H^{(2)}$ we performed second-order perturbation theory in the degenerate manifolds of H_{cost}, each of which contained independent sets of a fixed size. These manifolds become non-degenerate when exchanging H_{cost} for the Rydberg Hamiltonian,

$$H_{\text{Ryd}} = -\delta \sum_{u \in V} n_u + \sum_{u,v} V_{uv} n_u n_v, \quad \text{(C27)}$$

due to the long-range interactions $V_{uv} \sim 1/|r_u - r_v|^6$. At sufficiently large distances $|r_u - r_v|$, $V_{u,v}$ is small and has negligible effect. However, to safely perform perturbation theory in $(\Omega/\delta)_\star$, we must carefully handle the Rydberg interaction energy at short distances.

In the experimental implementation, the avoided level crossing occurs at a detuning of $\delta_\star \simeq 7$–$13\,\text{MHz}$. The energy scale for H_q is $|\Omega| = 2\,\text{MHz}$ (note that our definition of Ω differs from the standard definition of Rabi frequency by a factor of 2). Although the resulting value of $(\Omega/\delta)_\star \ll 1$, the necessary condition to perform perturbation theory is that the energy difference under H_{Ryd} between independent sets connected via H_q is large compared to Ω. Therefore, in addition to δ, we must consider the interaction energy V_{uv}, which is 107 MHz for nearest-neighbors on the square lattice and 13.6 MHz for next-nearest neighbors (see Fig. 1, main text, and Supplementary Information of Ref. [23]). Suppose we take an independent set and add a vertex via H_q, creating an independent set violation between nearest-neighbors.

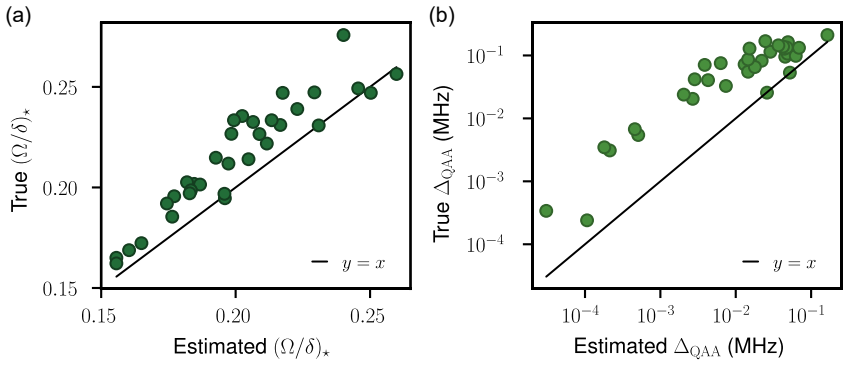

FIG. 9. Perturbation theory on the Rydberg Hamiltonian. (a) The true location of the avoided level crossing $(\Omega/\delta)_\star$ is close to the predicted value from perturbation theory, particularly for small $(\Omega/\delta)_\star$. Here we use Hamiltonian energy scales identical to those used in the experimental implementation. (b) The true minimum gap Δ_{QAA} can be estimated by only considering low-order terms in the resolvent formalism.

This interaction can be treated perturbatively because the energy difference between an independent set with and without a single nearest-neighbor violation under H_{Ryd} is $\geq 94\,\mathrm{MHz} \gg |\Omega|$. However, suppose we instead add a vertex that creates a single independent set violation with a next-nearest-neighbor. The new energy under H_{Ryd} increases by at least 13.6 MHz due to V_{uv}, and decreases by 13 MHz due to δ, meaning that this transition can be near-resonant under H_q. Therefore, we must treat single independent set violations between next-nearest neighbors non-perturbatively. We find that for most instances, removing a vertex from an independent set via a spin flip can be treated perturbatively, and discuss rare exceptions later.

We will use the standard Schrieffer-Wolff transformation to compute $H^{(2)}$ for the Rydberg Hamiltonian. By the above arguments, $|\Omega|$ is perturbatively small compared to the energy difference between near-degenerate manifolds of states that include:

1. Valid independent sets of the same size, and
2. Independent sets with any number of next-nearest neighbor independent set violations (where each vertex has at most a single next-nearest neighbor in the Rydberg state).

Of course, these configurations are not truly degenerate under H_{Ryd} due to long range interactions, but their splitting is comparable to $|\Omega|$, and typically small compared to the energy splitting between adjacent manifolds, which is approximately 13.6 MHz (up to interactions that are longer-range than next-nearest neighbors). Within a near-degenerate manifold, $H^{(2)}$ is given by

$$H^{(2)} = H_{\mathrm{Ryd}} - H_q - \frac{1}{2}[S, H_q], \tag{C28}$$

where the H_q term implicitly acts only within a near-degenerate manifold (i.e., it only (de)excites single next-nearest-neighbor independent set violations). The third term $-\frac{1}{2}[S, H_q]$ describes perturbative interactions between neighboring manifolds due to H_q, where S satisfies

$$-H_q + [S, H_{\mathrm{Ryd}}] = 0. \tag{C29}$$

Solving Eq. (C29) for S gives

$$\langle z|\, S\, |z''\rangle = -\frac{\langle z|\, H_q\, |z''\rangle}{H_{\mathrm{Ryd}}(z) - H_{\mathrm{Ryd}}(z'')}. \tag{C30}$$

Here z and z'' are in adjacent manifolds connected by H_q. Therefore we see explicitly that the perturbative condition is $H_{\mathrm{Ryd}}(z) - H_{\mathrm{Ryd}}(z'') \gg |\Omega|$.

Inserting S into Eq. (C29), we find that $H^{(2)}$ is given by

$$\begin{aligned} H^{(2)}_{z,z'} = {} & \langle z|\, H_{\mathrm{Ryd}}\, |z'\rangle - \langle z|\, H_q\, |z'\rangle \\ & + \sum_{z'':\langle z|H_q|z''\rangle\langle z''|H_q|z'\rangle \neq 0} \frac{\Omega^2}{2}\Big(\frac{1}{H_{\mathrm{Ryd}}(z) - H_{\mathrm{Ryd}}(z'')} \\ & + \frac{1}{H_{\mathrm{Ryd}}(z') - H_{\mathrm{Ryd}}(z'')}\Big), \end{aligned} \tag{C31}$$

for z, z' in the same near-degenerate manifold (here we have removed couplings involving two vertex additions or removals because they connect different manifolds, and are therefore off-resonant). Eq. (C31) is identical to the perturbative Hamiltonian for H_{QAA} [Eq. (11)], but with the denominator of the second-order terms replaced with the energy difference under H_{Ryd} instead of H_{cost}. We note that for a small number of instances, there exists one or more independent sets $|z\rangle$ such that removing a single vertex creates an independent set $|z''\rangle$ for which $H_{\mathrm{Ryd}}(z) - H_{\mathrm{Ryd}}(z'') \leq |\Omega|$, because the Rydberg interaction energy from the removed vertex is comparable to $-\delta$. We observe that this occurs only when the removed vertex cannot spin-exchange, so only the corresponding contribution to the second-order diagonal energy shift in $H^{(2)}$ is non-perturbative. In these rare cases we modify

this matrix entry to be the hybridized energy of $|z\rangle$ and $|z''\rangle$.

Given our expression for $H^{(2)}$, we can now compute the parameters involved in the avoided level crossing. For each graph instance, we enumerate the independent sets of size α and $\alpha-1$ using a tensor network algorithm [40], which is easily achieved on a laptop for the system sizes we study ($n = 39-80$). From the independent sets we construct $H^{(2)}$ and find its lowest energy eigenstate and eigenenergy for a given value of Ω/δ, which corresponds to the leading order approximation for $|\mathcal{G}\rangle$ or $|\mathcal{E}\rangle$ under S. From this, we can predict $(\Omega/\delta)_\star$ by finding the value of Ω/δ where the perturbed energies of $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ intersect. The energy where $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ intersect provides an estimate of $E_\star$. Figure 9(a) shows that the estimated $(\Omega/\delta)_\star$ from perturbation theory agrees with the true $(\Omega/\delta)_\star$ computed via DMRG, particularly as $(\Omega/\delta)_\star$ becomes small.

Using our perturbatively estimated $|\mathcal{G}\rangle, |\mathcal{E}\rangle$, and $(\Omega/\delta)_\star$, we can now estimate the minimum gap Δ_{QAA}. Ideally we would evaluate Eq. (8) in the main text, replacing H_{cost} with H_{Ryd}, but this is intractable at the largest system sizes we study ($n = 65, 80$). Inspired by the form of Eq. (8), we instead compute $\Delta^{\text{est.}}_{\text{QAA}}$, an estimate for Δ_{QAA} given by

$$\Delta^{\text{est.}}_{\text{QAA}} = 2\sum_{z,z'} (\Omega/\delta)_\star^{d(z,z')} \langle z|\mathcal{G}\rangle \langle z'|\mathcal{E}\rangle, \tag{C32}$$

where $d(z,z')$ is the pairwise Hamming distance between z and z'. $\Delta^{\text{est.}}_{\text{QAA}}$ corresponds to only considering the lowest-order coupling between $|z\rangle \in |\mathcal{G}\rangle$ and $|z'\rangle \in |\mathcal{E}\rangle$ under H_q in Eq. (8), which approximately occurs at order $(\Omega/\delta)_\star^{d(z,z')}$. In Fig. 9(b), we show that $\Delta^{\text{est.}}_{\text{QAA}}$ and Δ_{QAA} are similar. This verifies that even low-order estimations can qualitatively predict Δ_{QAA}. We note that $\Delta^{\text{est.}}_{\text{QAA}}$ can be computed with relatively low space complexity on the order of $\mathcal{O}(D_\alpha + D_{\alpha-1})$, where recall D_b is the number of independent sets of size b.

Appendix D: The star graph

Here we analyze the QAA runtime to find the largest independent set of a family of star graphs. A star graph has n_b branches of even length ℓ connected by a central vertex. We are interested in the runtime as a function of n_b at fixed ℓ.

1. Level-crossing parameters

We start by deriving the parameters involved in the avoided level crossing when $(\Omega/\delta)_\star \to 0$. In this limit, we can perturbatively predict $(\Omega/\delta)_\star$, the ground state energy at the avoided crossing $E_\star$, and the states involved in the avoided crossing $|\mathcal{G}\rangle, |\mathcal{E}\rangle$. We can determine these parameters from the eigenstates and eigenenergies of the second-order perturbed Hamiltonian (Eq. (11), main text),

$$H^{(2)} = -\frac{\Omega^2}{\delta}\left(H_{se} + \sum_{u\in V}\left[n_u - (\mathbb{1}-n_u)\prod_{(u,v)\in E}(\mathbb{1}-n_v)\right]\right).$$

$|\mathcal{G}\rangle$ is the ground state of $H^{(2)}$ in the manifold of independent sets with $H_{\text{cost}} = -\delta\alpha$. This corresponds to the unique largest independent set of the star graph with $\alpha = \frac{\ell n_b}{2} + 1$ vertices, including the central vertex and alternating vertices on each branch (see Fig. 3(c), main text). The eigenenergy of $|\mathcal{G}\rangle$ in $H^{(2)}$ corresponds to the second-order energy shift of $|\mathcal{G}\rangle$. It has nonzero contributions only from the second term of $H^{(2)}$, which evaluates to $-\frac{\Omega^2}{\delta}\sum_{u\in V} n_u = -\frac{\Omega^2}{\delta}\alpha$. The other terms are zero because no spin-exchange operations are possible and no vertices can be added to the largest independent set. Therefore at second-order the energy of $|\mathcal{G}\rangle$ is given by

$$\langle\mathcal{G}|H_{\text{cost}} + H^{(2)}|\mathcal{G}\rangle = -\delta\alpha - \frac{\Omega^2}{\delta}\alpha. \tag{D1}$$

Next we determine $|\mathcal{E}\rangle$ and its corresponding energy shift by finding the ground state of $H^{(2)}$ in the $H_{\text{cost}} = -\delta(\alpha-1)$ manifold. In this manifold there are $(\ell/2+1)^{n_b}$ independent sets of size $\alpha-1$ with the central vertex absent, and each branch in one of the $\ell/2+1$ largest independent sets of a one-dimensional length-ℓ chain with open boundary conditions (see Fig. 10(a), top). This degeneracy corresponds to the motion of a single domain wall (two adjacent vertices absent from the independent set) in the antiferromagnetic ordering on each branch. There are also a small number of independent sets of size $\alpha-1$ with the central vertex present (see Fig. 10(a), bottom). In these sets, all but one of the branches has perfect anti-ferromagnetic ordering ($\ell/2$ vertices in the set per branch), and the remaining branch has $\ell/2-1$ vertices. One can count that the number of such independent sets is $3n_b(\ell/2-1)$, meaning they form a vanishingly small fraction of independent sets of size $\alpha-1$ as $n_b \to \infty$.

As n_b grows we find that $|\mathcal{E}\rangle$ primarily has support on the independent sets with the central vertex absent. We first observe that the first term in $H^{(2)}$, $-\frac{\Omega^2}{\delta}H_{se}$, determines the ground state of $H^{(2)}$ to good approximation. To see this, first note that the second term in $H^{(2)}$ acts uniformly on all independent sets of size $\alpha-1$, so it does not affect the eigenvectors. The third term gives a small diagonal shift onto an independent set for every vertex that can be added to that set. This term is zero for all but $\ell n_b/2+1$ independent sets that connect to the largest independent set via a single spin flip, where it gives a shift of $\frac{\Omega^2}{\delta}$. When n_b is large, this energy shift is negligible compared to the ground state energy of the remaining term $-\frac{\Omega^2}{\delta}H_{se}$, which maximizes the expected number of spin exchanges. In particular, the ground state energy of this term is dominated by independent sets with the central vertex absent, which have anywhere between n_b and

$2n_b$ possible spin exchanges, depending on if the domain wall is on the boundary (one possible spin exchange) or in the bulk (two spin exchanges) of that branch. In comparison, independent sets with the central vertex present have only one or two total possible spin exchanges.

Therefore $|\mathcal{E}\rangle$ is well-approximated as the ground state of $-\frac{\Omega^2}{\delta}H_{se}$ restricted to the independent sets with the central vertex absent. On each branch this acts as a one-dimensional hopping Hamiltonian with open boundary conditions for the single domain wall. $|\mathcal{E}\rangle$ is therefore the product of the ground state over all n_b branches

$$\langle x_1 x_2 \dots x_{n_b}|\mathcal{E}\rangle = \prod_{i=1}^{n_b} \frac{1}{\sqrt{\ell/4+1}} \sin\left(\frac{\pi x_i}{\ell/2+2}\right), \quad \text{(D2)}$$

where $|x_i\rangle, x_i \in \{1,2,\dots,\ell/2+1\}$ is the state with the domain wall on the ith branch between sites $2x_i-2$ and $2x_i-1$. We confirm numerically that the overlap of Eq. (D2) with the true ground state of $H^{(2)}$ quickly approaches one as n_b grows for $\ell \in \{2,4,6,8\}$.

The corresponding perturbed energy at second-order is then

$$\begin{aligned}&\langle\mathcal{E}|H_{\text{cost}} + H^{(2)}|\mathcal{E}\rangle \\ &\quad\simeq -\delta(\alpha-1) - \frac{\Omega^2}{\delta}(\alpha-1) - \frac{\Omega^2}{\delta}\langle\mathcal{E}|H_{se}|\mathcal{E}\rangle. \quad \text{(D3)}\end{aligned}$$

By our earlier reasoning, $\langle\mathcal{E}|H_{se}|\mathcal{E}\rangle = c_\ell n_b$ where $c_\ell \in [1,2]$ is a computable number depending on ℓ. For $\ell = 2$, each configuration in $|\mathcal{E}\rangle$ can spin-exchange n_b times (once on each branch), so $c_\ell = 1$. As ℓ increases, the $|\mathcal{E}\rangle$ localizes on configurations that can spin-exchange $2n_b$ times (with the domain walls in the bulk of each branch), so $c_\ell \to 2$.

Having computed $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ and their energies as a function of Ω/δ, we can now estimate $(\Omega/\delta)_\star$ and $E_\star$ to second order in Ω/δ. $(\Omega/\delta)_\star$ is the value of Ω/δ where the two perturbed energies, Eqs. (D1) and (D3), intersect, given by

$$(\Omega/\delta)_\star = \sqrt{\frac{1}{c_\ell n_b - 1}}. \quad \text{(D4)}$$

This quantity goes to zero as $n_b \to \infty$, verifying that our perturbation theory converges as $n \to \infty$ at fixed ℓ. Fig. 10(b) shows the predicted and numerically computed (via exact diagonalization) value of $(\Omega/\delta)_\star$ for $\ell = 2$ and 4, which have $c_\ell = 1$ and $\sqrt{2}$, respectively. We reach system sizes of 100 and 97 for $\ell = 2$ and 4, respectively, by symmetrizing the Hamiltonian over the branches of the star graph.

The corresponding ground state energy is computed by evaluating Eq. (D1) at $(\Omega/\delta)_\star$. This gives, in units of δ,

$$-\frac{E_\star}{n} = \frac{\alpha}{n}\left(1 + \frac{1}{c_\ell n_b - 1}\right). \quad \text{(D5)}$$

Figure 10(c) shows the predicted and actual values of $-E_\star/n$ for the same instances at $\ell = 2, 4$. As expected, the predicted values converge to the true values as n_b increases.

2. Quantum runtime

We now compute the minimum gap $\Delta_{\text{QAA}}(\ell, n_b)$ and analyze its scaling as a function of n_b at fixed ℓ. We will show that $\Delta_{\text{QAA}}(\ell, n_b)$ scales as

$$\Delta_{\text{QAA}}(\ell, n_b) = \mathcal{O}\left(\Omega\left[\frac{1}{\sqrt{\ell/4+1}}\sin\left(\frac{\pi}{\ell/2+2}\right)\right]^{n_b}\right), \quad \text{(D6)}$$

up to polynomial factors in n_b, which are subleading compared to Eq. (D6), which is exponentially small in n_b. This matches the scaling predicted from leading-order perturbation theory in $(\Omega/\delta)_\star$ in Eq. (15) from the main text.

Following the resolvent formalism discussed in Appendix C 1, we will evaluate the estimated minimum gap $\tilde{\Delta}_{\text{QAA}}$. Recall from Eqs. (C3) and (C4) that $\tilde{\Delta}_{\text{QAA}}$ is given by the off-diagonal matrix element of an effective Hamiltonian $H_{\text{eff}}(z)$ acting on the subspace spanned by $|\mathcal{G}\rangle$ and $|\mathcal{E}\rangle$,

$$\begin{aligned}\tilde{\Delta}_{\text{QAA}} &= 2|\langle\mathcal{G}|H_{\text{eff}}(z)|\mathcal{E}\rangle| \quad \text{(D7)}\\ &= 2\left|\langle\mathcal{G}|H_{\text{cost}} - H_q + H_q\frac{Q}{z - QH_{\text{QAA}}Q}H_q|\mathcal{E}\rangle\right|,\end{aligned}$$

where $H_{\text{QAA}} = H_{\text{cost}} - H_q$ is evaluated at $(\Omega/\delta)_\star$, and $z \simeq E_\star$ is a parameter with dimensions of energy. By the resolvent formalism equation for the minimum gap in Eq. (C12), $\tilde{\Delta}_{\text{QAA}}(\ell, n_b)$ gives $\Delta_{\text{QAA}}(\ell, n_b)$ up to a computable proportionality factor that depends on $dH_{\text{eff}}(z)/dz$ and which is close to one. We numerically verify the correctness of Eq. (C12) in Fig. 10(d) by computing $\Delta_{\text{QAA}}(\ell, n_b)$ for $\ell = 2, 4$ via both exact diagonalization and by numerically evaluating Eq. (C12). To compute Eq. (C12), we first compute $H_{\text{eff}}(z)$, which gives us $\tilde{\Delta}_{\text{QAA}}$ by Eq. (D7). We compute the proportionality factor by evaluating $dH_{\text{eff}}(z)/dz$ using the finite difference method. When this correction factor is applied to $\tilde{\Delta}_{\text{QAA}}(\ell, n_b)$, the result matches $\Delta_{\text{QAA}}(\ell, n_b)$ computed via exact diagonalization to high accuracy, as expected. We observe numerically that $\tilde{\Delta}_{\text{QAA}} \simeq 4.53\Delta_{\text{QAA}}$ for $\ell = 2$, and $\tilde{\Delta}_{\text{QAA}} \simeq 7.85\Delta_{\text{QAA}}$ for $\ell = 4$, approximately independently of n_b. Therefore, as argued in Appendix C 1, $\tilde{\Delta}_{\text{QAA}}$ captures the relevant scaling of Δ_{QAA} in n_b.

To simplify the computation of Eq. (D6), we use a slightly different choice of $|\mathcal{E}\rangle$ from the previous Appendix D 1. This is allowed as long as $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ are reasonable approximations to the eigenstates involved in the avoided level crossing (see Appendix C 2). We let $H^{(i)}_{\text{QAA}}$ equal H_{QAA} restricted to the ith branch of the star graph. We let $|\mathcal{E}_i\rangle$ be the ground state of $H^{(i)}_{\text{QAA}}$, and choose $|\mathcal{E}\rangle = \otimes_{i=1}^{n_b}|\mathcal{E}_i\rangle$. Note that $|\mathcal{E}\rangle$ is equal to the ground state of H_{QAA} from second-order degenerate perturbation theory [Eq. (D2)], to leading order in $(\Omega/\delta)_\star$.

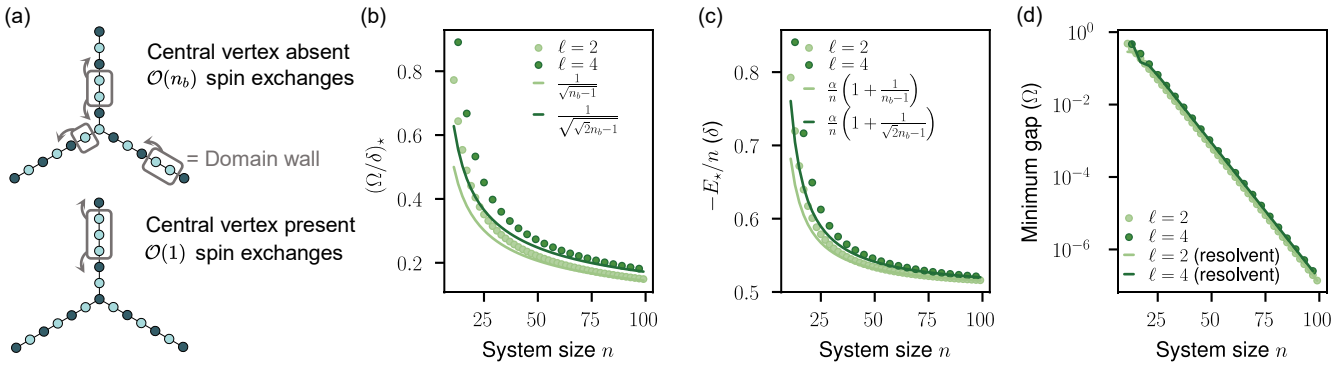

FIG. 10. Perturbative avoided level crossing in the star graph. (a) There are two types of suboptimal independent sets of size $\alpha - 1$ in the star graph (dark blue vertices are present in the independent set, light blue vertices are absent). Sets with the central vertex absent have a single domain wall on each of n_b branches that can hop to neighboring sites via spin exchanges, yielding $\mathcal{O}(n_b)$ possible spin exchanges per independent set. When the central vertex is present, only one branch has a domain wall, so there are $\mathcal{O}(1)$ possible spin exchanges. These latter independent sets have negligible amplitude in $|\mathcal{E}\rangle$, which favors independent sets with more possible spin exchanges. The predicted (solid lines) and numerically computed (data points) values of $(\Omega/\delta)_\star$ (b) and $-E_\star/n$ (c) match as n increases at fixed branch length ℓ. As $n \to \infty$, $(\Omega/\delta)_\star \to 0$. (d) The minimum gap computed via exact diagonalization matches the gap computed by numerically evaluating the resolvent method formula Eq. (C12) in Appendix C 1. At fixed ℓ, the minimum gap decreases exponentially as a function of n_b (and therefore n).

We now evaluate Eq. (D7). The first term yields $\langle\mathcal{G}| H_{\text{cost}} |\mathcal{E}\rangle = 0$. The second term is of the same order as Eq. (D6) because our $|\mathcal{E}\rangle$ is equal to the prediction from second-order degenerate perturbation theory to leading order in n_b (see Eq. (15), main text). Therefore it remains to compute the third term, $\langle\mathcal{G}| H_q \frac{Q}{z - QH_{\text{QAA}}Q} H_q |\mathcal{E}\rangle$. We begin by simplifying the outermost factors of H_q. First, we define the state

$$|\tilde{\mathcal{E}}_i\rangle = (\mathbb{1} - |\mathcal{E}_i\rangle\langle\mathcal{E}_i|)(\mathbb{1} - |\mathcal{G}\rangle\langle\mathcal{G}|) H_q^{(i)} |\mathcal{E}_i\rangle , \tag{D8}$$

where $H_q^{(i)}$ is H_q restricted to a single branch i. Then,

$$QH_q |\mathcal{E}\rangle = \sum_{i=1}^{n_b} |\tilde{\mathcal{E}}_i\rangle \otimes_{j\neq i} |\mathcal{E}_j\rangle , \tag{D9}$$

where we have used the fact that $H_q |\mathcal{E}\rangle$ is a sum of n_b terms, in each of which $n_b - 1$ branches are in $|\mathcal{E}_i\rangle$.

Meanwhile, when H_q acts on $|\mathcal{G}\rangle$ on the left hand side of the third term of Eq. (D7), one term in H_q removes the central vertex from $|\mathcal{G}\rangle$, yielding the state $\otimes_{i=1}^{n_b} |x_i = 1\rangle$, where $|x_i = 1\rangle$ denotes that the domain wall on the ith branch is on the first site (see Appendix D 1). $H_q |\mathcal{G}\rangle$ also contains terms in which vertices are removed from the branches of $|\mathcal{G}\rangle$, while the central vertex is left excited. These terms cannot have better scaling with n_b than the term with the central vertex removed from $|\mathcal{G}\rangle$, which we confirm numerically. They are higher order because to connect to $|\mathcal{G}\rangle$ via these terms, one must first go through $\otimes_{i=1}^{n_b} |x_i = 1\rangle$ to add the central vertex. Therefore, we have

$$\begin{aligned}&\langle\mathcal{G}| H_q \frac{Q}{z - QH_{\text{QAA}}Q} H_q |\mathcal{E}\rangle \\ &\quad = n_b \left(\otimes_{i=1}^{n_b} \langle x_i = 1|\right) Q \left[\frac{\mathbb{1}}{z - QH_{\text{QAA}}Q}\right] |\tilde{\mathcal{E}}_1\rangle \otimes_{i=2}^{n_b} |\mathcal{E}_i\rangle .\end{aligned} \tag{D10}$$

Here we have specified without loss of generality that the factor of $|\tilde{\mathcal{E}}_i\rangle$ occurs on $i = 1$, which yields the factor of n_b.

We will now make the approximation that Q factorizes between branches,

$$QH_{\text{QAA}}Q \approx \sum_{i=1}^{n_b} QH_{\text{QAA}}^{(i)}Q, \tag{D11}$$

where $H_{\text{QAA}}^{(i)}$ is H_{QAA} restricted to a single branch i. This is an approximation because it neglects terms in H_{QAA} which act on the central vertex of the graph. This leaves us with (up to polynomial factors in n_b)

$$\begin{aligned}&\langle\mathcal{G}| H_q \frac{Q}{z - QH_{\text{QAA}}Q} H_q |\mathcal{E}\rangle \sim \\ &\quad \left(\otimes_{i=1}^{n_b} \langle x_i = 1|\right) Q \frac{\mathbb{1}}{z - \sum_{i=1}^{n_b} QH_{\text{QAA}}^{(i)}Q} |\tilde{\mathcal{E}}_1\rangle \otimes_{i=2}^{n_b} |\mathcal{E}_i\rangle .\end{aligned} \tag{D12}$$

Note now that $[QH_{\text{QAA}}^{(i)}Q, QH_{\text{QAA}}^{(j)}Q] = 0$, so that we

may use the identity

$$\frac{\mathbb{1}}{z-\sum_{i=1}^{n_b} QH_{\mathrm{QAA}}^{(i)}Q} = \frac{1}{(2\pi i)^{n_b-1}}\int dz_1\dots dz_{n_b}\left[\delta\left(z-\sum_{i=1}^{n_b} z_i\right)\times\prod_{i=1}^{n_b}\frac{\mathbb{1}}{z_i-QH_{\mathrm{QAA}}^{(i)}Q}\right], \tag{D13}$$

where the z_i integrals are taken on a contour encircling the real axis. Therefore, we have

$$\langle\mathcal{G}|\,H_q\frac{Q}{z-QH_{\mathrm{QAA}}Q}H_q\,|\mathcal{E}\rangle \sim \frac{1}{(2\pi i)^{n_b-1}}\int dz_1\dots dz_{n_b}\left[\delta\left(z-\sum_{i=1}^{n_b} z_i\right)\times(\otimes_{i=1}^{n_b}\langle x_i=1|)\,Q\prod_{i=1}^{n_b}\frac{\mathbb{1}}{z_i-QH_{\mathrm{QAA}}^{(i)}Q}\,|\tilde{\mathcal{E}}_1\rangle\otimes_{i=2}^{n_b}|\mathcal{E}_i\rangle\right]. \tag{D14}$$

Note now that Q acts trivially on $|\tilde{\mathcal{E}}_1\rangle\otimes_{i=2}^{n_b}|\mathcal{E}_i\rangle$, because $|\tilde{\mathcal{E}}_1\rangle$ has no overlap with $|\mathcal{E}_1\rangle$. Furthermore, $H_{\mathrm{QAA}}^{(n_b)}$ only changes the state on branch n_b, so that we can write

$$\frac{\mathbb{1}}{z_{n_b}-QH_{\mathrm{QAA}}^{(n_b)}Q}\,|\tilde{\mathcal{E}}_1\rangle\otimes_{i=2}^{n_b}|\mathcal{E}_i\rangle = |\tilde{\mathcal{E}}_1\rangle(\otimes_{i=2}^{n_b-1}|\mathcal{E}_i\rangle)\frac{\mathbb{1}}{z_{n_b}-H_{\mathrm{QAA}}^{(n_b)}}\,|\mathcal{E}_{n_b}\rangle\,. \tag{D15}$$

We can repeat this process n_b-2 more times to obtain

$$\langle\mathcal{G}|\,H_q\frac{Q}{z-QH_{\mathrm{QAA}}Q}H_q\,|\mathcal{E}\rangle \sim \frac{1}{(2\pi i)^{n_b-1}}\int dz_1\dots dz_{n_b}\left[\delta\left(z-\sum_{i=1}^{n_b} z_i\right)(\otimes_{i=1}^{n_b}\langle x_i=1|)\times Q\frac{\mathbb{1}}{z_1-QH_{\mathrm{QAA}}^{(1)}Q}\left(|\tilde{\mathcal{E}}_1\rangle\otimes_{i=2}^{n_b}\frac{\mathbb{1}}{z_i-H_{\mathrm{QAA}}^{(i)}}|\mathcal{E}_i\rangle\right)\right]. \tag{D16}$$

We then make the replacement $\frac{\mathbb{1}}{z_i-H_{\mathrm{QAA}}^{(i)}}|\mathcal{E}_i\rangle\to 2\pi i\delta(z_i-\epsilon)\,|\mathcal{E}_i\rangle$, where ϵ is the energy of $|\mathcal{E}_i\rangle$ under H_{QAA}. This is valid by our choice of integration contour and because $|\mathcal{E}_i\rangle$ is an eigenstate of $H_{\mathrm{QAA}}^{(i)}$. Performing the z_i integrals yields

$$\langle\mathcal{G}|\,H_q\frac{Q}{z-QH_{\mathrm{QAA}}Q}H_q\,|\mathcal{E}\rangle \sim (\otimes_{i=1}^{n_b}\langle x_i=1|)\times Q\frac{\mathbb{1}}{z-(n_b-1)\epsilon-QH_{\mathrm{QAA}}^{(1)}Q}\left(|\tilde{\mathcal{E}}_1\rangle\otimes_{i=2}^{n_b}|\mathcal{E}_i\rangle\right). \tag{D17}$$

At this point, formally, the factors of $Q=\mathbb{1}-|\mathcal{G}\rangle\langle\mathcal{G}|-\prod_{i=1}^{n_b}|\mathcal{E}_i\rangle\langle\mathcal{E}_i|$ act on all factors in the wavefunction. However, since all but one of the n_b factors in the tensor product on the right are $|\mathcal{E}_i\rangle$, and since $H_{\mathrm{QAA}}^{(1)}$ only changes the state on the branch i which is not in $|\mathcal{E}_i\rangle$, we may safely replace Q with $Q_1=\mathbb{1}-|\mathcal{E}_1\rangle\langle\mathcal{E}_1|$, and obtain

$$\langle\mathcal{G}|\,H_q\frac{Q}{z-QH_{\mathrm{QAA}}Q}H_q\,|\mathcal{E}\rangle \sim (\otimes_{i=1}^{n_b}\langle x_i=1|)\times Q\frac{\mathbb{1}}{z-(n_b-1)\epsilon-Q_1H_{\mathrm{QAA}}^{(1)}Q_1}\left(|\tilde{\mathcal{E}}_1\rangle\otimes_{i=2}^{n_b}|\mathcal{E}_i\rangle\right). \tag{D18}$$

At this point, the final factor of Q may be dropped, because $\frac{\mathbb{1}}{z-(n_b-1)\epsilon-Q_1H_{\mathrm{QAA}}^{(1)}Q_1}|\tilde{\mathcal{E}}_1\rangle$ has no overlap with $|\mathcal{E}_1\rangle$. The expression becomes

$$\langle\mathcal{G}|\,H_q\frac{Q}{z-QH_{\mathrm{QAA}}Q}H_q\,|\mathcal{E}\rangle \sim (\langle x_1=1|\mathcal{E}_1\rangle)^{n_b-1}\times\langle x_1=1|\frac{\mathbb{1}}{z-(n_b-1)\epsilon-Q_1H_{\mathrm{QAA}}^{(1)}Q_1}|\tilde{\mathcal{E}}_1\rangle\,. \tag{D19}$$

The factor of $\langle x_1=1|\frac{\mathbb{1}}{z-(n_b-1)\epsilon-Q_1H_{\mathrm{QAA}}^{(1)}Q_1}|\tilde{\mathcal{E}}_1\rangle$ should scale at most polynomially with n_b and is thus subleading, by the arguments presented in Appendix C 1. The term $\langle x_1=1|\mathcal{E}_1\rangle)^{n_b-1}$ scales as Eq. (D6). Therefore, we conclude that $\tilde{\Delta}_{\mathrm{QAA}}$ (and therefore Δ_{QAA}) has the same asymptotic scaling with n_b as Eq. (D6).

Appendix E: Runtime of the modified QAA

In this section, we will analyze the optimized runtime $\Delta_{\mathrm{QAA}}^{-1}$ of the modified QAA [Eq. (16), main text]. We will show that $\Delta_{\mathrm{QAA}}^{-1}$ scales as the square root of the classical Markov chain runtime lower bounds from Appendix A under motivated assumptions about the energy landscape. We numerically verify this when the H_ℓ energy scale $\lambda\to\infty$ for system sizes of up to 460 vertices in Fig. 4(c) of the main text. Here, we provide an analytic arguments supporting these numerical observations. We first analyze the case where $\lambda\to\infty$ next in Appendix E 1. Perturbative corrections to our arguments for the case when λ is finite are discussed in Appendix E 2.

1. Infinite λ case

In the $\lambda\to\infty$ limit, the adiabatic dynamics are projected onto the ground subspace of H_ℓ. The ground subspace of H_ℓ is spanned by the uniform superpositions of each independent set size $\{|S_b\rangle\}_{b=0,\dots,\alpha}$ [Eq. (9), main text] when there exists a path between any two independent sets of the same size via spin exchanges. Here we assume this condition is met, and discuss exceptions in Appendix E 2. Each uniform superposition $|S_b\rangle$ experiences an energy shift of $-\delta b$ from H_{cost} and couples to neighboring independent set sizes under H_q with coupling strength $t_b=\langle S_b|\,H_q\,|S_{b-1}\rangle=\Omega b\sqrt{D_b/D_{b-1}}$. Therefore,

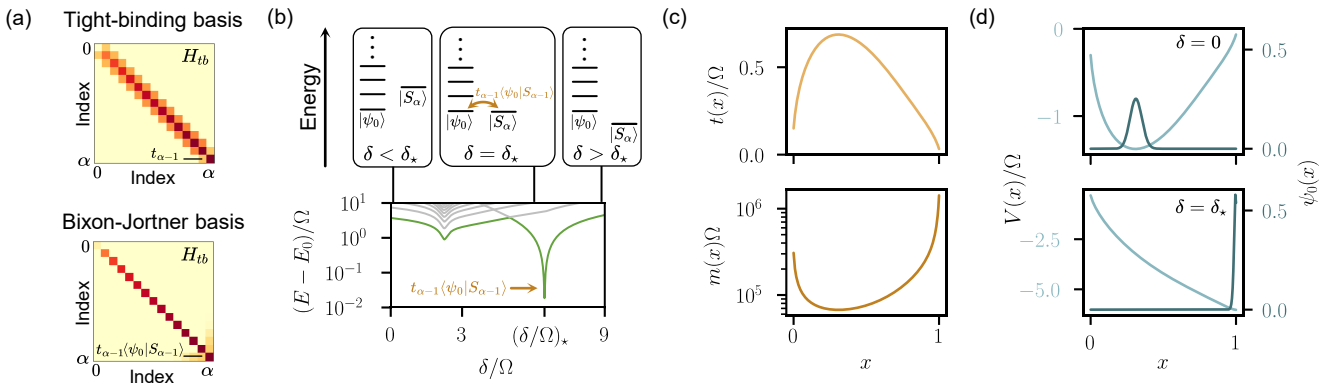

FIG. 11. Minimum gap of the modified QAA at infinite λ. (a) The original one-dimensional tight-binding Hamiltonian H_{tb} has a weak coupling $t_{\alpha-1}$ between the last and second-to-last sites (top). H_{tb} can be partially diagonalized to generate an effective Bixon-Jortner model that weakly couples all H_{bulk} eigenstates to the last site of the tight-binding model (bottom). (b) The lowest energy eigenvalues of H_{tb} as a function of δ/Ω for a representative instance with $n = 720$ vertices (bottom) are paired with schematic eigenenergies of H_{bulk} and the last site $|S_\alpha\rangle$ at three different detunings (top). For $\delta < \delta_\star$, the spectral gap of H_{tb} is equal to the spectral gap of H_{bulk}. At the resonance condition $\delta = \delta_\star$ between the last site and the H_{bulk} ground state $|\psi_0\rangle$, the weak coupling $t_{\alpha-1}$ sets the gap. For $\delta > \delta_\star$, Wannier-Stark localization sets in and the gap is proportional to δ. (c) We show a representative example of the couplings $t(x)$ and position dependent mass $m(x) = \frac{\alpha^2}{t(x)}$ from a 720-vertex hard unit-disk graph instance. (d) For the same instance, we show the effective potential $V(x)$, neglecting the second-derivative term $\partial_x^2 t(x)$, and the H_{bulk} ground state wavefunction $\psi_0(x)$ for $\delta = 0$ and $\delta = \delta_\star$. At $\delta = 0$, the ground state is delocalized in the middle of the bulk (top). At $\delta = \delta_\star$, the wavefunction is localized near the weak coupling (bottom).

the effective dynamics are given by the one-dimensional tight-binding Hamiltonian H_{tb} [Eq. (18), main text],

$$H_{tb} = -\sum_{b=1}^{\alpha} \left[\delta b \left|S_b\right\rangle \left\langle S_b\right| + t_b(\left|S_b\right\rangle \left\langle S_{b-1}\right| + \text{h.c.})\right]. \quad \text{(E1)}$$

Our goal is to show that the minimum gap Δ_{QAA} of H_{tb} goes like the smallest coupling $\min_b t_b$. For simplicity, we will focus on instances where the smallest coupling is between the largest independent sets of size α and suboptimal independent sets of size $\alpha - 1$, i.e. $\min_b t_b = t_{\alpha-1}$. This was overwhelmingly the most common case, representing 99.87% of the hundreds of instances studied in Appendix B. If $\Delta_{\text{QAA}} \propto t_{\alpha-1}$, then the modified optimized QAA runtime $\Delta_{\text{QAA}}^{-1} \propto t_{\alpha-1}^{-1}$ is quadratically smaller than the classical Markov chain runtime $\propto t_{\alpha-1}^{-2}$, up to polynomial factors in n. These polynomial factors are insignificant because numerically, $t_{\alpha-1}^{-2}$ is exponentially large in $\sqrt{n}$ for the Maximum Independent Set problem on unit-disk graphs (see Appendix B).

We first leverage the assumption that $t_{\alpha-1}$ is the smallest parameter in H_{tb}. We bipartition the system into two parts: the last site, corresponding to $|S_\alpha\rangle$, and the remaining sites which form the "bulk" of the chain. These two parts are connected by the weakest coupling $t_{\alpha-1}$:

$$H_{tb} = H_{\text{bulk}} - \delta\alpha \left|S_\alpha\right\rangle \left\langle S_\alpha\right| - t_{\alpha-1}(\left|S_\alpha\right\rangle \left\langle S_{\alpha-1}\right| + \text{h.c.}), \quad \text{(E2)}$$

where

$$H_{\text{bulk}} = -\sum_{b=0}^{\alpha-1} \left[\delta b \left|S_b\right\rangle \left\langle S_b\right| + t_b(\left|S_b\right\rangle \left\langle S_{b-1}\right| + \text{h.c.})\right]. \quad \text{(E3)}$$

We then diagonalize H_{bulk} and re-express H_{tb} in terms of its eigenenergies E_l and eigenvectors $|\psi_l\rangle$ (where $l = 0, \ldots, \alpha - 1$ is ordered from lowest to highest energy):

$$H_{tb} = -\delta\alpha \left|S_\alpha\right\rangle \left\langle S_\alpha\right| + \sum_{l=0}^{\alpha-1} \left[E_l \left|\psi_l\right\rangle \left\langle \psi_l\right| - t_{\alpha-1} \left\langle \psi_l|S_{\alpha-1}\right\rangle (\left|\psi_l\right\rangle \left\langle S_\alpha\right| + \text{h.c.})\right]. \quad \text{(E4)}$$

Eq. (E4) is a Bixon-Jortner model [70], a standard model in quantum optics where uncoupled levels interact with each other only by coupling to a common mode. Here, the uncoupled eigenstates $|\psi_l\rangle$ of H_{bulk} are each coupled to the last site of the tight-binding chain (the common mode) with strength $t_{\alpha-1} \langle\psi_l|S_{\alpha-1}\rangle$, as in Fig. 11(a). The coupling to the common mode is generated by projecting the last site onto the energy eigenstates of H_{bulk}: $\langle\psi_l| H_q |S_\alpha\rangle = t_{\alpha-1} \langle\psi_l| S_{\alpha-1}\rangle$.

We now consider what happens to the spectral gap of H_{tb} as we vary the detuning δ at fixed $\Omega = 1$, which we visualize in Fig. 11(b). We let $\delta_\star$ denote the detuning corresponding to when the ground state of H_{bulk}, $|\psi_0\rangle$, and the last site $|S_\alpha\rangle$ are resonant in energy (i.e., $E_0 = -\delta_\star\alpha$). From the canonical solution of the Bixon-Jortner problem [70], it follows that once $E_0, E_1, \cdots E_{\alpha-1} > -\delta_\star\alpha$, the spectral gap increases due to level repulsion as δ is

increased. In the language of the original tight-binding Hamiltonian, when $\delta > \delta_\star$, the electric field δ dominates so that the instance-specific details of the couplings t_b become irrelevant, and Wannier-Stark localization occurs in the bulk (see Fig. 11(b), top right). Therefore, the spectral gap is set by δ and the smallest coupling $t_{\alpha-1}$ does not play a role in determining the gap for $\delta > \delta_\star$. When $\delta < \delta_\star$, the spectral gap of H_{tb} is set by the spectral gap of H_{bulk}, which we denote as Δ_{bulk} (see Fig. 11(b), top left).

The gap is sensitive to the smallest coupling $t_{\alpha-1}$ at $\delta_\star$. Then, the ground state energy of H_{bulk}, $|\psi_0\rangle$, is resonant with the energy of the last site $|S_\alpha\rangle$, i.e., $E_0(\delta_\star) = -\delta_\star \alpha$. Here we show that the minimum gap Δ_{QAA} is given by the gap at the resonance,

$$\Delta_{\text{QAA}} = t_{\alpha-1} \langle \psi_0 | S_{\alpha-1} \rangle + \mathcal{O}(t_{\alpha-1}^2), \tag{E5}$$

when the following condition holds:

1. The spectral gap Δ_{bulk} of H_{bulk} is at least polynomially small in n for all values of δ.

This first condition guarantees two things: first, that two-level Landau-Zener physics occurs at $\delta = \delta_\star$, as the bulk ground state $|\psi_0\rangle$ and the last site $|S_\alpha\rangle$ are energetically well-separated from higher excited eigenstates of H_{tb}. This ensures that at $\delta = \delta_\star$, the gap is given by the Bixon-Jortner coupling $t_{\alpha-1} \langle \psi_0 | S_{\alpha-1} \rangle$. Second, it guarantees that the avoided level crossing at $\delta = \delta_\star$ is the *minimum* gap, as the gap for $\delta < \delta_\star$, Δ_{bulk}, is larger than Eq. (E5).

Δ_{QAA} is thus quadratically smaller than the classical Markov Chain runtime lower bounds up to polynomial factors in n when a second condition holds:

2. $\langle \psi_0 | S_{\alpha-1} \rangle$ is, at least, polynomially small in $1/n$.

This condition guarantees that the magnitude of the Bixon-Jortner matrix coupling $\Delta_{\text{QAA}} = t_{\alpha-1} \langle \psi_0 | S_{\alpha-1} \rangle$ at $\delta_\star$ is set by $t_{\alpha-1}$ and not by localization of $|\psi_0\rangle$ at sites other than $|S_{\alpha-1}\rangle$. Therefore, it is sufficient to show that $|\psi_0\rangle$ at $\delta_\star$ has at least polynomial in n overlap near the $(\alpha-1)$st site in the chain. If both of these conditions hold, Δ_{QAA}^{-1} is quadratically enhanced over the inverse of the classical Markov chain runtime up to polynomial factors in n.

We show next that both of these conditions are met under motivated assumptions about the couplings t_b. We numerically analyze hundreds of hard Maximum Independent Set instances on large graphs (from 460 to 720 vertices). Our numerical investigations of the couplings t_b reveal that while the specifics of the couplings vary from instance to instance, *in the bulk* they can, empirically, be well-described by a smooth function of the site index b along the tight-binding chain. Note that this condition often breaks down at the interface between the α and $\alpha-1$ because $t_{\alpha-1}$ is exponentially small in $\sqrt{n}$, but we have crucially split that term from the bulk. In passing we note that the normalized couplings, $t(x) \equiv t_b/\alpha$ appears to converge to a near-universal curve across hundreds of instances as a function of $x = b/\alpha$ for small to intermediate $0 < x < 0.5$, and as $\frac{1}{\sqrt{\alpha}}$ for small x (one can easily check this for the $x = 1/\alpha$ case). The normalized couplings peak at a constant value $\simeq 0.69$ before displaying instance-to-instance variation as they become small for $x \to 1$, as displayed in Fig. 12(a).

Motivated by these numerical observations, we now state constraints on the couplings that imply both conditions are satisfied.

Theorem 6. *Assume that the couplings t_b for $b = 0, \ldots, \alpha-1$ are a smooth, weakly concave function $t(x)$ of $x = \frac{b}{\alpha}$, that $t(1) \to 0$ as $\frac{1}{n^\gamma}$ for some $\gamma > 0$, and $\int_0^1 t(x)^{-1/2} dx$ is at most polynomially large in the system size n. Furthermore, assume $\delta_\star$ is sufficiently large such that $V(x) = -\delta x - 2t(x) + \frac{1}{\alpha^2}\partial_x^2 t(x)$ is locally minimized for $1 - \frac{1}{\alpha} < x < 1$. Then both conditions (1) and (2) hold.*

Proof. We appeal to the continuum limit of H_{bulk}, taken as the system size $n \to \infty$. This is equivalent to taking the largest independent set size $\alpha \to \infty$, since for the unit-disk graphs embedded on a two-dimensional lattice with constant filling fraction, α is proportional to n. We can take H_{bulk} to the continuum limit because we assumed that the couplings t_b are a smooth function of the site x. The new continuum, time-independent Schrodinger equation for the eigenstates $\psi(x)$ in the bulk is, for arbitrary δ,

$$\left(-\frac{1}{\alpha^2} \partial_x t(x) \partial_x + V(x) \right) \psi(x) = \varepsilon \psi(x), \tag{E6}$$

where $V(x) = -\delta x - 2t(x) + \frac{1}{\alpha^2}\partial_x^2 t(x)$ and ε is the energy density (energy normalized by α). Note that the site-dependent couplings have two major contributions. First, they induce a *position-dependent mass* going as $m(x) = \frac{\alpha^2}{t(x)}$, which imposes a metric on the chain. Second, the couplings induce a potential energy given by $-2t(x)$. The term that goes as the second derivative in the couplings is kept as it may grow with n when $t(x)$ goes to zero at the boundaries $x = 0, 1$. Away from the boundaries of the bulk, this second derivative term is negligible as $\alpha \to \infty$. In Fig. 11(c), we visualize $t(x)$ and $m(x)$ for an example unit-disk graph instance with $n = 720$ vertices. We plot the corresponding potential $V(x)$, neglecting the second derivative term, in Fig. 11(d) for $\delta = 0$ and $\delta = \delta_\star$.

We can arrive at a more conventional position-independent problem by performing two similarity transforms. The first is a point canonical transformation,

$$u(x) = \int_0^x \frac{1}{\sqrt{t(y)}} dy, \tag{E7}$$

which transforms the position dependent mass term into a position independent kinetic term: $\partial_x t(x) \partial_x \to \partial_u^2$. We then employ the standard integrating factor to remove

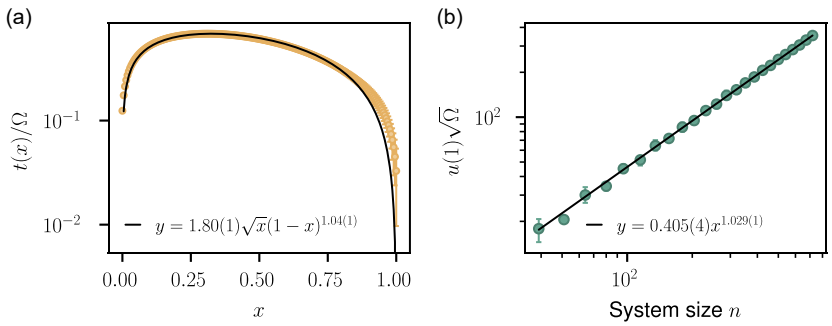

FIG. 12. Parameters of the one-dimensional tight-binding Hamiltonian. (a) We show the mean couplings $t(x)$ generated from 229 unique 720-vertex unit-disk graph instances with largest independent set size $\alpha = 217$. Error bars give the maximum and minimum coupling over all instances. $t(x)$ is well-described by a smooth function with universal behavior for $x \lesssim 0.3$. There are significant instance-by-instance variations in the couplings as $x \to 1$. We find that the mean couplings $t(x)$ are well-fit to the functional form $y = 1.80(1)\sqrt{x}(1-x)^{1.04(1)}$ (black line). (b) The mean value of $u(1) = \int_0^1 t(y)^{-1/2} dy$ grows polynomially with the system size n. As a result, the gap of H_{bulk} vanishes at most polynomially in $1/n$. Error bars give the maximum and minimum value of $u(1)$ over 1000 instances for each system size.

terms that are first order in the spatial derivative, leading to a typical Schrodinger problem: $\left(-\frac{1}{\alpha^2}\partial_u^2 + U(u)\right)\psi(u) = \varepsilon\psi(u)$, for a doubly-transformed effective potential $U(x)$. The similarity transformations do not alter the smoothness nor the convexity of the original potential. Moreover, the contributions from $\frac{1}{\alpha^2}\partial_x^2 t(x)$ do not change the convexity of the potential. As such, the effective potential U meets the weak convexity criterion stipulated in Andrews and Clutterbucks' proof of the fundamental gap conjecture for one-dimensional systems [71]. Therefore, we can apply the fundamental gap conjecture to bound Δ_{bulk} for all δ as

$$\Delta_{\text{bulk}} \geq \frac{3\pi^2}{(\alpha u(1))^2}. \tag{E8}$$

Thus, as long as the couplings can be well-described by a smooth $t(x)$ and $u(1)$ grows at most polynomially in n, Δ_{bulk} is polynomially small in $1/n$. This proves our first condition.

To validate our second condition – that the ground state of the bulk is localized around the penultimate site on the chain – we provide a semi-classical argument. Note that the semi-classical approximation is well-justified in the limit of large system sizes $\alpha \to \infty$ as the effective mass $m(x) = \frac{\alpha^2}{t(x)}$, diverges at the edges (equivalently, $t(x) \to 0$). The resonance condition implies that the ground state energy density of the bulk, ε, at the crossing is $-\delta_\star$. The classical minimum of the potential approaches the edge of the chain in the regime of $\delta = \delta_\star$. As shown in Fig. 11(d), for $\delta = \delta_\star$, in order to minimize energy, the particle seeks to lower its potential due to the electric field gradient versus the potential due to tunneling. As the semi-classical expectation becomes exact in the limit of an infinite mass, the particle localizes near the edge where the classical minimum of the potential lies. Thus, within an asymptotically exact semi-classical argument, the ground-state of the bulk should localize near the site corresponding to $|S_{\alpha-1}\rangle$. This validates the second condition of our argument. As a result, we have shown that in the $\lambda \to \infty$ limit, Δ_{QAA}^{-1} is quadratically smaller than the classical Markov chain runtime lower bounds.

a. *Numerical justification*

By examining 1000 unit-disk Maximum Independent Set problem instances for each system size between 460 and 720 vertices, in Fig. 12(a) we find a simple qualitative model for $t(x)$ is given by $t(x) = Ax^{\frac{1}{2}}(1-x)^c$. The factor of $\frac{1}{2}$ encodes the exact scaling as $x \to 0$ and the fit parameter c accounts for instance-to-instance variations as $x \to 1$. We find that the fitted values of c fall between $c \in \{1.03, 1.09\}$ across different instances, leading to appropriate conditions for $t(x)$ such that $u(1)$ is only polynomially large in n. We confirm numerically that $u(1)$ grows approximately linearly with n in Fig. 12(b). Thus, our numerical results confirm that Δ_{bulk} is at worst polynomially small in $1/n$, under the assumption of the validity of our continuum analysis. Therefore, it is well-justified to focus on the resonant level crossing between the last site and the ground state of the bulk only to determine Δ_{QAA}.

This numerical evidence also supports our assumption about $V(x)$ being locally minimized for $1 - \frac{1}{\alpha} < x < 1$. One might worry that the diverging mass near the edges of the tight-binding model causes the bulk ground state wavefunction to be classically forbidden from penetrating the region of the penultimate site on the chain. Indeed, by solving Eq. (E6) for the ground state with energy density $-\delta$ and using the WKB approximation, one notices that there are two classically forbidden regimes: from smaller x due to the an increase in the potential, and at $x \to 1$ due to terms originating from the diverging mass

(e.g. terms proportional to $[\partial_x^2 t(x)]/t(x)$). However, the latter classically forbidden regime, following the qualitative model for $t(x) = Ax^{1/2}(1-x)^c$, occurs for $x > 1-\frac{c}{2\alpha}$, where c is numerically fitted to be within 1.03 and 1.09. This suggests that the classically forbidden regime occurs within the penultimate site, which occupies $1-\frac{1}{\alpha} < x < 1$ on the continuum, which can be seen by simply inverting the mapping from the continuum back to discrete sites on a chain. Thus our numerics also strongly suggest that the wavefunction is localized around $|S_{\alpha-1}\rangle$, such that $\langle\psi_0|S_{\alpha-1}\rangle$ is sufficiently large and $\Delta_{\mathrm{QAA}} = t_{\alpha-1}$, up to polynomial factors in n.

b. *Optimizing the modified QAA*

For QAA to achieve a runtime which scales as $\Delta_{\mathrm{QAA}}^{-1}$ in practice, the algorithm schedule $(\Omega(t), \delta(t))$ must be optimized so that its parameters change slowly near the location of the avoided level crossing (see Fig. 3(a), main text). In particular, by choosing $|dH/dt| \propto \Delta_{\mathrm{QAA}}^2$ within a $\mathcal{O}(\Delta_{\mathrm{QAA}})$ interval around the location of the minimum gap, the total QAA runtime is $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-1})$ [11]. It is therefore useful to be able to estimate the location of the avoided crossing, so that the algorithm schedule can be optimized. Techniques to optimize QAA are a subject of active research, and recent work indicates that it is possible to optimize QAA on a wide class of disordered cost Hamiltonians when using the reflection about the uniform superposition state, $\mathbb{1} - \frac{1}{2^n}\sum_{z,z'}|z\rangle\langle z'|$, to drive the evolution, instead of H_q [41]. Here we describe a simple way to optimize the modified QAA when $\lambda \to \infty$, which achieves a total runtime, including optimization, that is asymptotically smaller than the SA runtime $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-2})$. The full quadratic speedup, with runtime $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-1})$, is recovered if quantum phase estimation is used as a subroutine in optimizing the QAA. A partial speedup, with runtime $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-10/7})$, is obtained if only projective measurements in the σ_z and σ_x bases are used. We leave the optimization of the modified QAA at arbtitrary λ as a subject of future research, possibly by generalizing the results of Ref. [41].

Our arguments follow from Appendix E 1, whose results we summarize here. When $\lambda \to \infty$, the QAA system Hamiltonian simplifies to an effective one-dimensional tight-binding Hamiltonian H_{tb} [Eq. (18)]. The sites of H_{tb} correspond to the uniform superpositions of each independent set size, $\{|S_b\rangle\}$ $(b = 0, 1, \ldots, \alpha)$. Suppose the final coupling between $|S_{\alpha-1}\rangle$ and $|S_\alpha\rangle$, equal to $\Omega\alpha\sqrt{D_\alpha/D_{\alpha-1}}$, is small compared to all other couplings. Then, this coupling can be treated perturbatively, and the system is described by a Bixon-Jortner model [70]. Let us take $\Omega = 1$ and consider the system ground state as a function of δ, as visualized in Fig 11(b). At $\delta < \delta_\star$, where $\delta_\star$ denotes the location of the avoided crossing, the system ground state is the ground state $|\psi_0\rangle$ of a restricted Hamiltonian H_{bulk}, which includes all sites up to $|S_{\alpha-1}\rangle$. The avoided level crossing occurs when the energy $E_0(\delta)$ of $|\psi_0\rangle$ is resonant with the energy $-\delta\alpha$ of last site of the chain, $|S_\alpha\rangle$. Thus, the avoided crossing occurs when $E_0(\delta_\star) = -\delta_\star\alpha$ to high $\mathcal{O}(\Delta_{\mathrm{QAA}}^2)$ accuracy by the arguments of Appendix E 1. For $\delta > \delta_\star$, $|\psi_0\rangle$ is the first excited state of H_{tb}.

Therefore, if one can estimate ground state energy E_0 of H_{bulk}, and compare its value to the resonance condition $E_0(\delta_\star) = -\delta_\star\alpha$, one can estimate $\delta_\star$. Because QAA can prepare $|\psi_0\rangle$ efficiently, finding E_0 is computationally simple. Note that this is *distinct* from generically finding the ground state of the system Hamiltonian H_{tb}. In particular, suppose we run QAA for time T with a linear schedule for $\delta(t)$, stopping the evolution at the desired value of δ. The precise choice of T is algorithm-dependent and discussed below, but we always choose $1/T$ to be much less than $1/\Delta_{\mathrm{bulk}}^2$, where Δ_{bulk} is the minimum gap of H_{bulk}, but larger than Δ_{QAA}. Because $T^{-1/2}$ is small compared to the energy difference between the first and second excited state, this schedule should remain adiabatic with respect to all but the smallest gap Δ_{QAA}. This schedule is still highly diabatic with respect to Δ_{QAA}, however, for which an instantaneous ramp speed $\propto \Delta_{\mathrm{QAA}}^2$ is needed to maintain adiabaticity. When the QAA evolution is stopped at $\delta < \delta_\star$, the state thus prepared by QAA will therefore have high overlap with the ground state of H_{tb} (and thus H_{bulk}), while for $\delta > \delta_\star$ it will have high overlap with the first excited state of H_{tb} (thus the ground state of H_{bulk}). In particular, for all values of δ, because of the chosen ramp time, the prepared wavefunction will have unit amplitude in $|\psi_0\rangle$, up to small corrections from all other instantaneous eigenstates of H_{tb}, which contribute small errors to the energy of the state.

Therefore $|\psi_0\rangle$ can be prepared using QAA, and one can measure its energy with error ε in time $\varepsilon^{-\gamma}$. The value of γ depends on the method used to compute the energy: $\gamma = 1$ using quantum phase estimation [72], and $\gamma = 5/2$ using projective measurements in the σ_z and σ_x bases. The value of $5/2$ is the combined result of shot noise and the time required for a single QAA run. In particular, with a N runs of time T each, we expect shot noise at the level of $\mathcal{O}(N^{-1/2})$ and nonadiabatic corrections to E_0 at the level of $\mathcal{O}(T^{-2})$ [73]. To make both of these $\mathcal{O}(\varepsilon)$ one can choose $T = \varepsilon^{-1/2}, N = \varepsilon^{-2}$, for a total time of $\varepsilon^{-5/2}$. Because quantum phase estimation does not require repeated runs, we can simply choose $T = \Delta_{\mathrm{QAA}}^{-1}$ so that nonadiabatic corrections of order $T^{-2} = \mathcal{O}(\Delta_{\mathrm{QAA}}^2)$ are negligible, and still retain $\gamma = 1$.

Our procedure in Algorithm 1 thus uses binary search to efficiently find $\delta_\star$ by minimizing the absolute value of the prepared energy of $|\psi_0\rangle$ minus $-\delta\alpha$. In particular, we use this procedure to estimate $\delta_\star$ to some finite, high accuracy depending on γ. We find that it is optimal to estimate $\delta_\star$ to $\mathcal{O}(\Delta_{\mathrm{QAA}}^{2/(1+\gamma)})$ accuracy. We then run the modified QAA using an optimized schedule with runtime $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-1})$, as in Ref. [11], for candidate guesses

Algorithm 1: Optimizing the modified QAA

Data: A subroutine that estimates $E_0(\delta)$ with error ε in time $\varepsilon^{-\gamma}$, for $\delta < \delta_\star$. An initial guess r for the ratio $D_{\alpha-1}/D_\alpha$, and a scale factor $k > 1$ with which we will increase r by during each iteration of the optimization.

while *An independent set of size α has not been found using the modified QAA.* **do**

  $r \leftarrow kr$

  Use subroutine to constrain $\delta_\star$ to an $\mathcal{O}(r^{-1/(1+\gamma)})$ interval in time $\mathcal{O}(r^{\gamma/(1+\gamma)})$. Draw $r^{1/2-1/(1+\gamma)}$ regularly spaced guesses for $\delta_\star$ from this interval.

  for *each guess for $\delta_\star$ in the $\mathcal{O}(r^{-1/(1+\gamma)})$ interval, in $\mathcal{O}(r^{-1/2})$ increments* **do**

    Run the modified QAA in time $\mathcal{O}(r^{1/2})$ using a schedule that slows down at the current guess of $\delta_\star$, such that $|dH/dt| \propto 1/r$ in an $\mathcal{O}(r^{-1/2})$ range of the guess for $\delta_\star$.

  end

end

of $\delta_\star$ within this range of possible values. By using a grid search for $\delta_\star$, the optimal solution can be found with a speedup for any $\gamma > 0$. The total runtime of Algorithm 1 is $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-2\gamma/(1+\gamma)})$, which results in a speedup over SA for any $\gamma > 0$, because the SA runtime is $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-2})$. This runtime is the result of a compromise between measurement time, which improves the precision with which $\delta_\star$ is known, and the time spent grid searching for $\delta_\star$ using QAA with an optimized schedule. In particular, if a time $\Delta_{\mathrm{QAA}}^{-2\gamma/(1+\gamma)}$ is spent measuring $\delta_\star$ to accuracy $\mathcal{O}(\Delta_{\mathrm{QAA}}^{2/(1+\gamma)})$, one wins a factor of $\mathcal{O}(\Delta_{\mathrm{QAA}}^{2/(1+\gamma)})$ in runtime relative to the $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-2})$ time that is required when grid searching for $\delta_\star$ with *no* knowledge of $\delta_\star$. As a result, a total time of $\mathcal{O}(\Delta_{\mathrm{QAA}}^{-2\gamma/(1+\gamma)})$ is also spent grid searching for $\delta_\star$.

Note that in practice, one does not know Δ_{QAA} *a priori*, and must therefore search for this as well. This is done efficiently in Algorithm 1 through a grid search on an exponentially spaced grid. Finally, we note that the two methods discussed above (phase estimation and projective measurements) are only suggestions for the subroutine required by Algorithm 1. Any method (quantum or classical) which can estimate $E_0(\delta)$ to error ε in time $\varepsilon^{-\gamma}$ would suffice.

2. Finite λ case

a. *Numerical observations*

Here we extend the arguments for a quadratic speedup in Appendix E 1 to the case where λ is finite. We first numerically compare the runtime of the modified QAA at finite λ and the unmodified QAA ($\lambda = 0$) for the top 1% hardest instances of each system size, up to $n = 80$. To compute Δ_{QAA} for each instance and setting of λ, we use the ITensor implementation [50] of DMRG to find matrix product state representations of the ground and first excited state with bond dimension of up to 1500. We consider the system converged to its true ground state $|\psi_0\rangle$ once the truncation error falls below a threshold value of 10^{-8}. In practice, this criterion is typically satisfied after $\mathcal{O}(10^2)$ sweeps. Once $|\psi_0\rangle$ is obtained, we compute the first excited state by repeating this procedure but with the Hamiltonian $H' = H + V|\psi_0\rangle\langle\psi_0|$, where $V = 10$ is an energy penalty that ensures that the ground state of H' has negligible overlap with $|\psi_0\rangle$. We then minimize the corresponding energy gap between the ground and first excited state over Ω/δ to obtain Δ_{QAA}, using a large energy penalty $U = 100$ on independent set violations [see Eq. (1)].

We display the numerical results in Fig. 13(a). We observe that for $\lambda = 5$, $\Delta_{\mathrm{QAA}}^{-1}$ is proportional to the square root of the SA runtime lower bound (the light blue data points are parallel to the line $y = \sqrt{x}$). Furthermore, setting $\lambda = 1$ is sufficient to obtain a speedup on the vast majority of instances (medium blue data points). In both cases, the modified QAA vastly outperforms the unmodified QAA ($\lambda = 0$, dark blue points). The fact that the unmodified QAA does not frequently outperform SA suggests that typical instances of the unmodified QAA do not have favorable localization or delocalization in the ground and first excited eigenstates at the avoided level crossing, which would ensure a speedup over SA. Thus, the modification to QAA appears crucial to obtain a speedup over SA on these instances.

To support these numerical observations, in the following section we further obtain analytic conditions that are sufficient, albeit not necessary, to guarantee the quadratic speedup, up to subleading polynomial factors in n. As in Appendix E 1, we focus on instances where the smallest coupling is between the largest independent sets of size α and suboptimal independent sets of size $\alpha - 1$, which was overwhelmingly the most common case for the instances studied in Appendix B. We then show that when $\Delta_{\mathrm{QAA}} \simeq \Omega\alpha\sqrt{D_\alpha/D_{\alpha-1}}$ in the $\lambda \to \infty$ case, the same conditions hold for for finite $\lambda/\Omega, \lambda/\delta \gtrsim \Delta_{\ell,\alpha}^{-1}, \Delta_{\ell,\alpha-1}^{-1}$, where $\Delta_{\ell,b}$ is the spectral gap of the Laplacian Hamiltonian H_ℓ when restricted to independent sets of size b. To obtain the speedup in practice, it is necessary that the scaling advantage is maintained when the Hamiltonian energy scales are normalized in units of λ. By dividing the energy scales of the Hamiltonian by λ, one can see that this is equivalent to the condition that $\lambda\Delta_{\mathrm{QAA}}^{-1}$ is quadratically smaller than the classical Markov chain runtime lower bounds, up to subleading polynomial factors in n, where Δ_{QAA} is the minimum gap in units of Ω. Thus, the speedup is obtained when $\Delta_{\ell,\alpha}^{-1}, \Delta_{\ell,\alpha-1}^{-1}$ grow at most polynomially in n. In practice, we find that $\Delta_{\ell,\alpha-1}^{-1} \geq \Delta_{\ell,\alpha}^{-1}$, so $\Delta_{\ell,\alpha-1}$ determines the strength of λ sufficient for ensuring delocalization.

Figure 13(b) shows the scaling of $\Delta_{\ell,\alpha-1}$ as a function

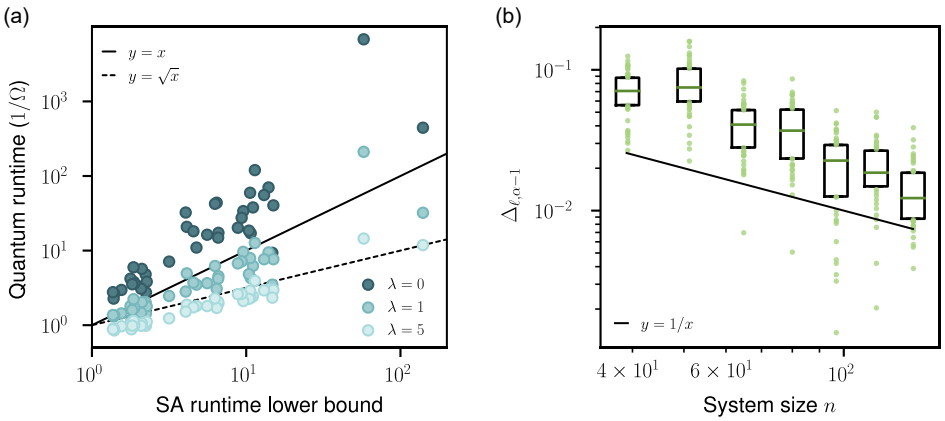

FIG. 13. Modified QAA runtime at finite λ. (a) The modified QAA runtime for $\lambda = 5$ scales as the square root of the SA runtime for the top 1% hardest instances of each system size up to $n = 80$ (light blue points). The speedup is also observed for the vast majority of instances for $\lambda = 1$ (medium blue points). In both cases, the modified QAA significantly outperforms the unmodified QAA ($\lambda = 0$, dark blue points). (b) For the top 5% hardest instances of each system size up to $n = 135$, $\Delta_{\ell,\alpha-1}$ is generally larger than $1/n$. The box endpoints mark the 25th and 75th percentiles, and the box midpoint marks the 50th percentile. We omit data for eight instances for which computing $\Delta_{\ell,\alpha-1}$ was too computationally expensive.

of n for the top 5% hardest instances up to $n = 135$. We observe that $\Delta_{\ell,\alpha-1} \gtrsim 1/n$ for the vast majority of instances, consistent with polynomial scaling in n. A minority of instances (24%) have $\Delta_{\ell,\alpha-1} = 0$ due to a very small fraction (median 0.2%) of configurations disconnected by spin exchanges, leading to degenerate ground states of H_ℓ in the manifold of independent sets of size $\alpha - 1$. For these instances, we plot the spectral gap of H_ℓ in the same manifold, restricted to the largest set of configurations connected under spin exchanges. One can see using perturbation theory that the smaller set(s) of disconnected configurations do not change the dynamics significantly, and the larger set determines the minimum gap. At small Ω/δ, the QAA Hamiltonian will energetically favor the connected subspace with the smaller expectation value of $-(\Omega^2/\delta)H_{se}$ under perturbation theory. This corresponds to the larger connected subspace, because the number of disconnected configurations in practice is very small (and thus, so is its expectation in H_{se}, which is upper-bounded by the maximum degree of a vertex in the configuration graph). We emphasize that the numerical results in Fig. 13(a) show that in practice, much smaller values of λ may be necessary to obtain the speedup, depending on the graph instance. All instances obtain a speedup for either $\lambda = 1$ or $\lambda = 5$, which is smaller than $\Delta^{-1}_{\ell,\alpha-1}$. This shows that while our condition is sufficient to obtain the speedup, it is not necessary in general.

Finally, it is interesting to note the connection between the gap $\Delta_{\ell,b-1}$ and the time needed for SA to sample from the equilibrium Gibbs distribution, restricted to a manifold of independent sets of the same size. This corresponds to SA sampling uniformly among independent sets of the same size. Consider an SA algorithm that only uses spin-exchange updates to explore independent sets of the same size (of course, this SA algorithm is only ergodic among independent sets of the same size, assuming all configurations can be connected via spin exchanges). One can check that H_ℓ is identical to the transition matrix used by SA, up to an overall rescaling and multiple of the identity. Thus, $\Delta_{\ell,b}$ sets the mixing time for SA to sample from the uniform distribution in that manifold. This idea can be generalized: consider an SA algorithm now using both spin-exchange and spin-flip updates. Again, the matrices within a manifold are identical up to rescaling when restricted to maximal independent sets (independent sets to which no vertices can be added without removing an existing vertex). The fraction of maximal independent sets is approximated by the quantity $1 - \frac{nD_b}{D_{b-1}}$, which is close to one on instances with a large SA runtime lower bound. Thus, H_ℓ and the SA transition matrix are near-identical, up to rescaling. As a result, $\Delta^{-1}_{\ell,b}$ sets the equilibration time to uniformly sample independent sets within that manifold. SA will thus need $\mathcal{O}(\Delta^{-1}_{\ell,\alpha-1})$ updates to converge to uniformly sample independent sets for the $\alpha - 1$ manifold. Because this quantity is polynomial in n, SA rapidly mixes within the $\alpha - 1$ manifold. The same is true of the manifold of independent sets of size α, because $\Delta^{-1}_{\ell,\alpha-1} \geq \Delta^{-1}_{\ell,\alpha}$. Thus, we expect the SA runtime $\tau_{\mathrm{SA}}(\varepsilon)$ is set by the time to find an optimal solution, which is exponential in $\sqrt{n}$, rather than the time to equilibrate within a manifold of independent sets of the same size. This is consistent with the numerical results in Fig. 2, which put together, suggests that $\tau_{\mathrm{SA}}(\varepsilon)$ is a good proxy for the SA time to find an optimal solution.

b. Sufficient analytic conditions for the speedup

We now show that $\Delta_{\mathrm{QAA}} \simeq \Omega\alpha\sqrt{D_\alpha/D_{\alpha-1}}$ for finite $\lambda/\delta, \lambda/\Omega \gtrsim \Delta_{\ell,\alpha}^{-1}, \Delta_{\ell,\alpha-1}^{-1}$. To this end, we will use the resolvent formalism developed in Appendix C 1, and let $|\mathcal{G}\rangle = |S_\alpha\rangle, |\mathcal{E}\rangle = |S_{\alpha-1}\rangle$. When $\lambda/\delta, \lambda/\Omega \gtrsim \Delta_{\ell,\alpha}^{-1}, \Delta_{\ell,\alpha-1}^{-1}$, we expect these states to have significant overlap with the ground and first-excited state of H_{QAA} at $(\Omega/\delta)_\star$. As a result,

$$\begin{aligned}\tilde{\Delta}_{\mathrm{QAA}} &= 2|\langle\mathcal{E}|H_{\mathrm{eff}}(E_\star)|\mathcal{G}\rangle| \\ &= 2\left|-\langle\mathcal{E}|H_q|\mathcal{G}\rangle + \langle\mathcal{E}|H_qQ\frac{Q}{E_\star - QHQ}QH_q|\mathcal{G}\rangle\right|\end{aligned} \tag{E9}$$

is a good estimator of Δ_{QAA} (see Appendix C 2), where $H = H_{\mathrm{cost}} - H_q + \lambda H_\ell$ is the modified QAA Hamiltonian. The first term of this expression is the coupling $-\Omega\alpha\sqrt{D_\alpha/D_{\alpha-1}}$ from the $\lambda \to \infty$ limit, which is responsible for the quadratic speedup. To argue that the speedup is maintained at finite λ, it remains to argue that the second term does not cancel with the first to reduce the gap. We do this by analyzing the dependence of this second term on λ.

We first simplify the second term. Let $|\tilde{\mathcal{G}}\rangle = \frac{1}{\alpha\Omega}H_q|\mathcal{G}\rangle$, where the $\alpha\Omega$ factor is used to make $\langle\tilde{\mathcal{G}}|\tilde{\mathcal{G}}\rangle \simeq 1$. Note that $|\tilde{\mathcal{G}}\rangle$ has support only on independent sets of size $\alpha-1$. We now write the second term from Eq. (E9) as

$$\begin{aligned}\langle\mathcal{E}|H_qQ\frac{Q}{E_\star - QHQ}QH_q|\mathcal{G}\rangle &= \alpha\Omega\langle\mathcal{E}|H_qQ\frac{Q}{E_\star - QHQ}Q|\tilde{\mathcal{G}}\rangle \\ &= \alpha\Omega\langle\mathcal{E}|H_qQ\frac{Q}{E_\star - QHQ}QH_qQ\frac{Q}{E_\star - Q(H_{\mathrm{cost}}+H_\ell)Q}Q|\tilde{\mathcal{G}}\rangle \\ &= \alpha\Omega\langle\mathcal{E}|H_qQ\frac{Q}{E_\star - QHQ}QH_qQ\frac{Q}{E_\star + \delta(\alpha-1) - QH_\ell Q}Q|\tilde{\mathcal{G}}\rangle\end{aligned} \tag{E10}$$

where in the second line we used the Woodbury matrix identity, and dropped a term which is unable to connect $|\tilde{\mathcal{G}}\rangle$ to $|\mathcal{E}\rangle$. Now, because we have taken $\lambda\Delta_{\ell,\alpha-1} \gg \Omega, \delta$, we may make the approximation

$$\frac{Q}{E_\star + \delta(\alpha-1) + QH_\ell Q}Q|\tilde{\mathcal{G}}\rangle \approx H_\ell^+|\tilde{\mathcal{G}}\rangle, \tag{E11}$$

where H_ℓ^+ denotes the Moore-Penrose pseudoinverse of H_ℓ, restricted to the space of sets of size $\alpha-1$. Here we rely on the fact that $E_\star + \delta(\alpha-1) = \mathcal{O}(\delta)$, as argued in Appendix C 3 because the avoided level crossing happens at $(\Omega/\delta)_\star \ll 1$. If the perturbative avoided level crossing condition is not met, then the same conclusion holds if we take $\lambda/\delta, \lambda/\Omega \gtrsim n\Delta_{\ell,\alpha}^{-1}, n\Delta_{\ell,\alpha-1}^{-1}$, which introduces a subleading factor of n to the runtime. We note that this approximation neglects a term that is $\mathcal{O}(\Omega/[\lambda\Delta_{\ell,\alpha-1}], \delta/[\lambda\Delta_{\ell,\alpha-1}])$, which we will argue below is subleading. The second term of Eq. (E9) thus reduces to

$$\begin{aligned}&\langle\mathcal{E}|H_qQ\frac{Q}{E_\star - QHQ}QH_q|\mathcal{G}\rangle \\ &\quad= \alpha\Omega\langle\mathcal{E}|H_qQ\frac{Q}{E_\star - QHQ}QH_qH_\ell^+|\tilde{\mathcal{G}}\rangle.\end{aligned} \tag{E12}$$

The central point of our argument is that the factor of H_ℓ^+, which scales as $\mathcal{O}(1/\lambda)$, ensures that the leading λ-dependence of this expression is $\mathcal{O}(1/\lambda)$. This will make it impossible for the second term of Eq. (E9) to always cancel exponentially with the first term. To see this, suppose for the sake of contradiction that for a specific value of λ, the second term was equal to $\alpha\Omega\sqrt{D_\alpha/D_{\alpha-1}}(1+\varepsilon)$, for some exponentially small ε (leading to a suppressed gap in Eq. (E9) of order $\alpha\Omega\sqrt{D_\alpha/D_{\alpha-1}}\varepsilon$). Then, if Eq. (E12) is $\mathcal{O}(1/\lambda)$, doubling λ will yield a gap from Eq. (E9) equal to $\alpha\Omega\sqrt{D_\alpha/D_{\alpha-1}}(1/2 - \varepsilon/2)$, which still achieves the quadratic speedup, losing only a factor of two.

It therefore remains to argue that Eq. (E12) decreases with λ as $1/\lambda$ or faster. Since H_ℓ^+ scales as $1/\lambda$, the only way this could not be the case is if the λ-dependence of the denominator $E_\star - QHQ$ changes the scaling to be slower than $1/\lambda$. This would occur if there were a leading order $\mathcal{O}(1/\lambda)$ term in $E_\star - QHQ$. However, since we have chosen $|\mathcal{G}\rangle, |\mathcal{E}\rangle$ to satisfy the overlap condition of Theorem 5 in Appendix C, we know that the smallest eigenvalue of $QHQ - E_\star$ is at least $(E_2 - E_\star)$, up to polynomial factors in $1/n$, where E_2 is the energy of the second excited state of H at the gap closing. In the limit we are considering, by standard perturbation theory in $\Omega, \delta/\lambda$, the leading (in particular, zeroth-order) contribution to $E_2 - E_\star$ will be independent of λ. As a result, the leading contribution to Eq. (E12) will be $\mathcal{O}(1/\lambda)$.

It is now also clear why the $\mathcal{O}(\Omega/[\lambda\Delta_{\ell,\alpha-1}], \delta/[\lambda\Delta_{\ell,\alpha-1}])$ term we dropped is unimportant. By nearly identical arguments to the above, this term will have a leading $\mathcal{O}(1/\lambda^2)$ scaling, which will not modify the overall argument that the second term in Eq. (E9) cannot cancel with the first term for generic values of λ (due to the λ-dependence of the second term).

[1] S. Arora and B. Barak, *Computational complexity: A modern approach* (Cambridge University Press, 2016).
[2] A. Montanaro, NPJ Quantum Inf. **2**, 15023 (2016).
[3] T. Albash and D. A. Lidar, Rev. Mod. Phys. **90**, 015002 (2018).
[4] D. J. Earl and M. W. Deem, Phys. Chem. Chem. Phys. **7**, 3910 (2005).
[5] E. Farhi, J. Goldstone, S. Gutmann, J. Lapan, A. Lundgren, and D. Preda, Science **292**, 472 (2001).
[6] A. Lucas, Front. Phys. **2** (2014).
[7] A. P. Young, S. Knysh, and V. N. Smelyanskiy, Phys. Rev. Lett. **104**, 020502 (2010).
[8] M. Guidetti and A. P. Young, Phys. Rev. E **84** (2011).
[9] I. Hen and A. P. Young, Phys. Rev. E **84** (2011).
[10] E. Farhi, J. Goldstone, and S. Gutmann, arXiv:quant-ph/0201031 [quant-ph] (2002).
[11] J. Roland and N. J. Cerf, Phys. Rev. A **65** (2002).
[12] S. Muthukrishnan, T. Albash, and D. A. Lidar, Phys. Rev. X **6**, 031010 (2016).
[13] E. Crosson and A. W. Harrow, in *57th Ann. IEEE Sympm. on Founds. of Comp. Sci. (FOCS)* (2016) pp. 714–723.
[14] M. Szegedy, in *45th Ann. IEEE Sympm. on Founds. of Comp. Sci. (FOCS)* (2004) pp. 32–41.
[15] R. D. Somma, S. Boixo, H. Barnum, and E. Knill, Phys. Rev. Lett. **101** (2008).
[16] A. Montanaro, Proc. R. Soc. A: Math. Phys. Eng. Sci. **471**, 20150301 (2015).
[17] M. W. Johnson, M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, E. M. Chapple, C. Enderud, J. P. Hilton, K. Karimi, E. Ladizinsky, N. Ladizinsky, T. Oh, I. Perminov, C. Rich, M. C. Thom, E. Tolkacheva, C. J. S. Truncik, S. Uchaikin, J. Wang, B. Wilson, and G. Rose, Nature **473**, 194 (2011).
[18] S. W. Shin, G. Smith, J. A. Smolin, and U. Vazirani, arXiv:1401.7087 [quant-ph] (2014).
[19] T. F. Rønnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, and M. Troyer, Science **345**, 420 (2014).
[20] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, Nat. Phys. **10**, 218 (2014).
[21] H. G. Katzgraber, F. Hamze, Z. Zhu, A. J. Ochoa, and H. Munoz-Bauza, Phys. Rev. X **5**, 031026 (2015).
[22] S. Boixo, V. N. Smelyanskiy, A. Shabani, S. V. Isakov, M. Dykman, V. S. Denchev, M. H. Amin, A. Y. Smirnov, M. Mohseni, and H. Neven, Nat. Commun. **7**, 10327 (2016).
[23] S. Ebadi, A. Keesling, M. Cain, T. T. Wang, H. Levine, D. Bluvstein, G. Semeghini, A. Omran, J.-G. Liu, R. Samajdar, X.-Z. Luo, B. Nash, X. Gao, B. Barak, E. Farhi, S. Sachdev, N. Gemelke, L. Zhou, S. Choi, H. Pichler, S.-T. Wang, M. Greiner, V. Vuletić , and M. D. Lukin, Science **376**, 1209 (2022).
[24] V. S. Denchev, S. Boixo, S. V. Isakov, N. Ding, R. Babbush, V. Smelyanskiy, J. Martinis, and H. Neven, Phys. Rev. X **6**, 031015 (2016).
[25] B. Altshuler, H. Krovi, and J. Roland, PNAS **107**, 12446 (2010).
[26] S. Lamm, P. Sanders, C. Schulz, D. Strash, and R. F. Werneck, J. Heuristics **23**, 207 (2017).
[27] D. Guéry-Odelin, A. Ruschhaupt, A. Kiely, E. Torrontegui, S. Martínez-Garaot, and J. Muga, Rev. Mod. Phys. **91** (2019).
[28] E. J. Crosson and D. A. Lidar, Nat. Rev. Phys. **3**, 466 (2021).
[29] A. D. King, J. Raymond, T. Lanting, R. Harris, A. Zucca, F. Altomare, A. J. Berkley, K. Boothby, S. Ejtemaee, C. Enderud, E. Hoskinson, S. Huang, E. Ladizinsky, A. J. R. MacDonald, G. Marsden, R. Molavi, T. Oh, G. Poulin-Lamarre, M. Reis, C. Rich, Y. Sato, N. Tsai, M. Volkmann, J. D. Whittaker, J. Yao, A. W. Sandvik, and M. H. Amin, Nature **617**, 61 (2023).
[30] B. F. Schiffer, D. S. Wild, N. Maskara, M. Cain, M. D. Lukin, and R. Samajdar, arXiv:2306.13131 [quant-ph] (2023).
[31] B. Clark, C. Colbourn, and D. Johnson, Discrete Math. **86**, 165 (1990).
[32] H. Pichler, S.-T. Wang, L. Zhou, S. Choi, and M. D. Lukin, arXiv:1808.10816 [quant-ph] (2018).
[33] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[34] W. K. Hastings, Biometrika **57**, 97 (1970).
[35] M. Žnidarič, Phys. Rev. A **71** (2005).
[36] The error $\varepsilon < 1/2$ is the total variation distance (equal to half the l_1 distance) between the Gibbs distribution and the distribution prepared by SA.
[37] D. A. Levin, Y. Peres, E. L. Wilmer, J. G. Propp, and D. B. Wilson, *Markov chains and mixing times* (American Mathematical Society, 2017).
[38] P. Diaconis and D. Stroock, The Annals of Applied Probability **1**, 36 (1991).
[39] The SA and QMC runtime lower bounds in Eqs. (5) and (19) assume that the independence polynomial of the graph instance is unimodal, i.e., $D_0 \leq D_1 \leq \cdots \leq D_{b^\star} \geq \cdots \geq D_{\alpha-1} \geq D_\alpha$ for some $b^\star$. In Appendix B, we study 24,000 unit disk graph instances with system sizes of up to 720 vertices, and find that every instance has a unimodal independence polynomial, which may be of independent interest [67]. If the independence polynomial is not unimodal, then the same bounds apply, but the maximum must be taken over all $b > b^\star$, as in Appendix A.
[40] J.-G. Liu, X. Gao, M. Cain, M. D. Lukin, and S.-T. Wang, SIAM J. Sci. Comput. **45**, A1239 (2023).
[41] M. Jarret, B. Lackey, A. Liu, and K. Wan, arXiv:1810.04686 [quant-ph] (2019).
[42] Nonperturbative calculation of transition amplitudes, in *Atom-Photon Interactions* (John Wiley & Sons, Ltd, 1998) Chap. 3, pp. 165–255.
[43] M. H. S. Amin, Phys. Rev. Lett. **100** (2008).
[44] M. H. S. Amin and V. Choi, Phys. Rev. A **80**, 062326 (2009).
[45] V. Choi, Quantum Inf. Process. **19**, 90 (2020).
[46] F. R. K. Chung, *Spectral graph theory*, Vol. 92 (CBMS Regional Conference Series in Mathematics, 1997).
[47] N. G. Dickson and M. H. Amin, Phys. Rev. A **85**, 032303 (2012).
[48] T. Lanting, A. D. King, B. Evert, and E. Hoskinson, Phys. Rev. A **96**, 042322 (2017).

[49] S. R. White, Phys. Rev. Lett. **69**, 2863 (1992).
[50] M. Fishman, S. White, and E. Stoudenmire, SciPost Phys. Codebases (2022).
[51] S. V. Isakov, G. Mazzola, V. N. Smelyanskiy, Z. Jiang, S. Boixo, H. Neven, and M. Troyer, Phys. Rev. Lett. **117** (2016).
[52] We include the factor of M in the definition of $\tau_{\mathrm{QMC}}(\varepsilon)$ to reflect the space-time complexity of a single QMC update. Because we allow our QMC update rule to modify all M Trotter slices, this complexity of $\mathcal{O}(M)$. Our results are unchanged if the normalization factor on $\tau_{\mathrm{QMC}}(\varepsilon)$ is taken to be n/m instead of n/M, where m is the number of Trotter slices modified during a single update.
[53] D. S. França and R. García-Patrón, Nat. Phys. **17**, 1221 (2021).
[54] D. Bluvstein, H. Levine, G. Semeghini, T. T. Wang, S. Ebadi, M. Kalinowski, A. Keesling, N. Maskara, H. Pichler, M. Greiner, V. Vuletić, and M. D. Lukin, Nature **604**, 451 (2022).
[55] A. Browaeys, D. Barredo, and T. Lahaye, J. Phys. B **49**, 152001 (2016).
[56] A. Browaeys and T. Lahaye, Nat. Phys. **16**, 132 (2020).
[57] M.-T. Nguyen, J.-G. Liu, J. Wurtz, M. D. Lukin, S.-T. Wang, and H. Pichler, Phys. Rev. X Quantum **4**, 010316 (2023).
[58] H. Krovi, M. Ozols, and J. Roland, Phys. Rev. A **82**, 022333 (2010).
[59] M. B. Hastings, Quantum **5**, 597 (2021).
[60] A. Gilyén and U. Vazirani, arXiv:2011.09495 [quant-ph] (2020).
[61] D. Gamarnik, PNAS **118**, e2108492118 (2021).
[62] E. Farhi, D. Gamarnik, and S. Gutmann, arXiv:2005.08747 [quant-ph] (2020).
[63] H. Bernien, S. Schwartz, A. Keesling, H. Levine, A. Omran, H. Pichler, S. Choi, A. S. Zibrov, M. Endres, M. Greiner, V. Vuletić, and M. D. Lukin, Nature **551**, 579 (2017).
[64] M. B. Hastings, Quantum **3**, 201 (2019).
[65] Y. Peres and P. Sousi, arXiv:1108.0133 [math.PR] (2013).
[66] J. Houdayer, Eur. Phys. J. B **22**, 479 (2001).
[67] V. E. Levit and E. Mandrescu, in *Proc. of the 1st Int. Conf. on Algebraic Inform.* (2005).
[68] F. V. Fomin and P. Kaski, Commun. ACM **56**, 80–88 (2013).
[69] M. Werner, A. García-Sáez, and M. P. Estarellas, arXiv:2301.13861 [quant-ph] (2023).
[70] E. J. Heller, *The Semiclassical Way to Dynamics and Spectroscopy* (Princeton University Press, 2018).
[71] B. Andrews and J. Clutterbuck, J. Am. Math. Soc. **24** (2011).
[72] A. Y. Kitaev, arXiv:quant-ph/9511026 [quant-ph] (1995).
[73] C. D. Grandi and A. Polkovnikov, in *Quantum Quenching, Annealing and Computation* (Springer Berlin Heidelberg, 2010) pp. 75–114.