Distance-Independent Entanglement Generation in a Quantum Network using Space-Time Multiplexed Greenberger–Horne–Zeilinger (GHZ) Measurements

Ashlesha Patil †
Wyant College of Optical Sciences
University of Arizona, Tucson AZ USA
ashlesha@email.arizona.edu

Joshua I. Jacobson †
Department of Physics,
University of Arizona, Tucson AZ USA
jjacob1@email.arizona.edu

Don Towsley
College of Information and Computer Sciences,
University of Massachusetts, Amherst MA USA
towsley@cs.umass.edu

Emily Van Milligen
Wyant College of Optical Sciences
University of Arizona, Tucson AZ USA
evanmilligen@email.arizona.edu

Saikat Guha
Wyant College of Optical Sciences
University of Arizona, Tucson AZ USA
saikat@optics.arizona.edu

Abstract—In a quantum network that successfully creates links—shared Bell states between neighboring repeater nodes—with probability $p$ in each time slot, and performs Bell State Measurements at nodes with success probability $q < 1$, the end-to-end entanglement generation rate drops exponentially with the distance between consumers, despite multi-path routing. If repeaters can perform multi-qubit projective measurements in the GHZ basis that succeed with probability $q$, the rate does not change with distance in a certain $(p, q)$ region, but decays exponentially outside. This region where the distance-independent rate occurs is the super-critical region of a new percolation problem. We extend this GHZ protocol to incorporate a time-multiplexing blocklength $k$, the number of time slots over which a repeater can mix-and-match successful links to perform fusion on. As $k$ increases, the super-critical region expands. For a given $(p, q)$, the entanglement rate initially increases with $k$, and once inside the super-critical region for a high enough $k$, it decays as $1/k$ GHZ states per time slot. When memory coherence time exponentially distributed with mean $\mu$ is incorporated, it is seen that increasing $k$ does not indefinitely increase the super-critical region; it has a hard $\mu$-dependent limit. Finally, we find that incorporating space-division multiplexing, i.e., running the above protocol independently in up to $d$ disconnected network regions, where $d$ is the network’s node degree, one can go beyond the $1$ GHZ state per time slot rate that the above randomized local-link-state protocol cannot surpass. As $(p, q)$ increases, one can approach the ultimate min-cut entanglement-generation capacity of $d$ GHZ states per slot.

Index Terms—Quantum networks, entanglement routing, time-multiplexing, GHZ projections, percolation

I. INTRODUCTION

Quantum networks are designed to distribute shared entangled states, which serve as an important resource for applications such as quantum key distribution (QKD), entanglement enhanced quantum sensing, and distributed quantum computing. Quantum networks are composed of fundamental building blocks called quantum repeaters, which are special purpose quantum computers that can store qubits (entangled, e.g., with qubits at neighboring repeaters) in a local quantum memory register, and perform quantum logic and measurements on sets of locally-held qubits. It is well known that, without the assistance of quantum repeaters, the shared entanglement generation rate drops exponentially with the distance between communicating parties [1], [2]. Several attempts have been made to improve the rate-versus-distance scaling starting with ideal as well as noisy Bell states generated across network links, followed by measurements performed at repeater nodes using local link state information, aimed at lengthening the reach of entanglement as well as to purify entanglement across links or multiple hops. In this paper, we will be concerned with quantum switches, but will continue to call them a repeater. A quantum switch is a repeater that has the capability of choosing which links to fuse, based on instantaneous network state information, dynamically.

With ideal two-qubit maximally-entangled Bell states generated with probability $p$ across each network in each time slot and repeaters allowed to perform Bell State Measurements (BSMs) that succeed with probability $q$, [3] showed that the path diversity (multi-path routing) afforded by the network improves the exponent of the exponential rate-distance scaling over what is achievable when the network routes entanglement over a pre-determined path. However, the rate still decays exponentially even with global link state knowledge as long as $q < 1$. Ref. [4] showed that time multiplexing over a linear chain of repeaters, i.e., allowing a repeater node to wait for blocks of $k$ time slots before deciding the pairs of qubits (within that time block) to be attempted to be fused (using BSMs), with an optimally-chosen $k$ can achieve a sub-exponential entanglement rate-vs.-distance scaling.
For the case when link-level Bell states have sub-unity fidelity, ref. [5] designed a quantum repeater architecture that uses multiplexing and a photonic switchboard to maximize the tradeoff between the shared entanglement rate and the fidelity of final entangled state. Ref. [6] on the other hand optimized the length of time that quantum memories should hold onto noisy Bell states to maximize the rate. Ref. [7] performed an extensive numerical optimization of the scheduling of purification and BSMs to maximize the distillable entanglement rate, given a target fidelity for the final shared entanglement state. 

Following the intuition of [4], ref. [8] evaluates the benefits of time-multiplexing in a two-dimensional quantum network. This paper extends the multi-path entanglement routing protocol presented in ref. [3] by allowing quantum memories to hold onto entangled qubits for multiple time-steps before deciding which qubit pairs to perform BSMs on. In the absence of quantum memory decoherence, time multiplexing can only help bridge the gap between the rate obtained using local link state knowledge and that obtained using global link state knowledge. When the BSM success probability \( q < 1 \), as the multiplexing block length increases, the obtainable rate monotonically increases. When memory decoherence is accounted for, the rate increases until some optimal time-multiplexing block length is reached, and then proceeds to decrease. In both cases there are regions of the parameter space \((p, q, k)\) where this dynamic local-link state knowledge based routing protocol can outperform simply routing over the predetermined shortest paths between end-consumers.

All of the above papers achieve rates that decrease with the distance between the communicating parties. In [9] however, it was shown that if repeaters are allowed to perform GHZ joint projective measurements (on 3 or more qubits), then in a certain \((p, q)\) region—determined by the critical threshold of a modified site-bond percolation problem—the entanglement rate does not decay with the end-to-end distance. This paper aims to improve the maximum entanglement of the protocol described in [9] by employing time-multiplexed GHZ projections, as well as spatial-division multiplexing. Our intent is to quantify the regime of the parameter space, i.e., \( p \), the probability that a network edge establishes a (noiseless) Bell state across it successfully in each time slot, \( q \), the probability that an attempted GHZ measurement succeeds, and \( k \), the time-multiplexing block length, which affords distance-independent entanglement rate. As the network has to wait for \( k \) time steps before the entangled state is shared between the consumers, the entanglement rate is defined as no. of shared entangled states generated per time slot and not ebits/s.

We employ two techniques to realize these improvements: (a) time-multiplexing to attain distance-independent rate for poorer \((p, q)\) parameters, and (b) spatial-division multiplexing to increase the maximum achievable rate. We also study the performance of the time-multiplexed protocol for the case that quantum memories are non-ideal and have finite coherence times that are exponentially-distributed with a known mean. This paper serves as a guidebook to determine the degree of time-multiplexing needed to achieve distance independent rate, but not use too high a \( k \) once inside the percolation region so as not to be faced with the \( 1/k \) rate penalty measured in end-to-end entangled states generated per time slot, & decide if spatial multiplexing is needed, given the knowledge of \( p \) and \( q \).

Finally, all mention of 'qubit' in this paper should be interpreted as being logical (error-corrected) qubits, and not physical qubits. This is because loss on a logical qubit can be heralded with the error-detection and error-correction property of a quantum code. This is an important assumption in modeling assumptions inherent in this paper. More discussion on this aspect will appear in Section III.

The organization of the paper is as follows. We start with a discussion of the time-multiplexed version of the GHZ measurement based protocol in Section II-A followed by the percolation problem the protocol translates to in Section II-B. The simulation results for the protocol are analysed in Section II-C. We relax the ideal quantum memory assumption made in Section II-A to discuss how the finite memory coherence times affect the rates and the optimal degree time-multiplexing in Section III. We discuss the spatially multiplexed version of the protocol in Section IV where we subdivide the network into spatially-disconnected regions while employing our protocol independently within each region, to achieve entanglement rate that approaches the ultimate capacity, the min-cut rate [10], for the underlying network topology. We conclude by summarising the main results in Section V. The code used to generate the simulation results is available at [11].

II. TIME-MULTIPLICED GHZ-PROJECTION BASED ENTANGLEMENT ROUTING

In this section, we introduce our time-multiplexed protocol where nodes perform up to \( n \)-qubit GHZ measurements, with qubits chosen over \( k \) time slots. We term this the \((n, k)\)-GHZ protocol. We describe the protocol, present simulated entanglement rates for a square grid network, followed by a new percolation problem our protocol maps to.

Let us begin with a few definitions. An \( n \)-qubit maximally entangled state of the form \(|0\rangle^{\otimes n} + |1\rangle^{\otimes n} \sqrt{\frac{2}{2}}\) is referred to as the \( n \)-GHZ state, which includes the Bell state \(|00\rangle + |11\rangle\sqrt{\frac{2}{2}}\), as a special case \((n = 2)\). There are \( 2^n \) mutually-orthogonal \( n \)-GHZ states, forming an orthonormal basis for the Hilbert space spanned by \( n \) qubits. Two fundamental quantum operations that are used in our entanglement-routing protocol, are as follows:

1. \( n \)-fusion—This is a probabilistic operation, which when successful, implements a joint projective measurement on \( n \) qubits (held at a repeater node) in the aforementioned \( n \)-GHZ basis. We assume that the \( n \)-fusion measurement projects the \( n \) measured qubits on to one of the \( 2^n \) GHZ states, \(|0\rangle^{\otimes n} + |1\rangle^{\otimes n} \sqrt{\frac{2}{2}}\). This does not result in a loss of generality, since if the \( n \)-fusion happens to project the \( n \) qubits on to any of the other \( 2^n - 1 \) \( n \)-GHZ states, one can apply local single-qubit Pauli unitary operations to bring the post-measurement state to the one obtained when the fusion projects on to the aforesaid standard
light. From the time when the entanglement attempt is initiated, to arrive at the nodes sharing that edge. Here, $\tau_q$ is the speed of light. We refer to a shared Bell pair, $|\phi^+\rangle = \frac{1}{\sqrt{2}} (|00\rangle + |11\rangle)$, between two neighboring repeater nodes as a link. In each time slot, of length $\tau$ seconds, an attempt is made to create a link across every network edge; it succeeds with probability $p$. The two qubits of a link are stored in quantum memory registers at the two nodes sharing the edge, if successful. The knowledge of the success-failure of each link takes time $\tau_L = L/c$ seconds, from the time when the entanglement attempt is initiated, to arrive at the nodes sharing that edge. Here, $L$ (meters) is the physical distance of the network edge, and $c$ is the speed of light.

We define the time-multiplexing block length, $k$, as the number of time-steps a repeater waits—in accumulating qubits entangled with neighboring nodes—before attempting fusions on a subset of the qubits accumulated, whose success-failure outcomes have arrived at the node. Fusions are destructive measurements. Performing an $n$-fusion thus frees up $n$ qubit slots in the memory register at the node. It should therefore be evident that at any given point in time, each repeater node holds between $d(\lceil \tau_L/\tau \rceil)$ and $d(\lceil \tau_L/\tau \rceil + k)$ qubits, where $d$ is the node’s degree ($d = 4$ for a square-grid network).

Every repeater can perform one- and two-qubit unitary operations and Pauli basis measurements on qubits held in its memory register. Fig. 1 depicts a $k = 3$ link-snapshot over a square-grid network. The line segments denote successful link attempts over a 3 time slot block. The blue dots represent qubits. In this section, we assume that the quantum memory registers are ideal in the sense that the qubit coherence time is much larger than $\tau_L + (k - 1)\tau$ seconds.

In our $(n,k)$-GHZ protocol, depicted schematically in Fig. 1, at the end of each block of $k$ time slots, nodes other than Alice or Bob, the communicating parties (the ‘consumers’), perform fusions on groups of (up to $n$) qubits, to expand the reach of the shared entanglement across the network. The rules for picking the fusion groups are described below. Each repeater acts based on local link state knowledge, i.e., the link success-failure outcomes on its neighboring edges. Fusion outcomes are independent events, each with probability of success $q$.

Let us use $d$ to denote node degree. For a square-grid network, $d = 4$. We will evaluate the performance of the $(n,k)$-GHZ protocol described below, on the square-grid network, for $n = 3$ and $n = 4$.

- A repeater randomly picks one successful link across neighboring edges, over one $k$ time slot block, into groups of (at least 2 and at most $\min(d,n)$) qubits to attempt fusions on. It repeats the process until there are no more links across two (or more) distinct neighbors. All links within the $k$-slot block are treated as equal in forming fusion groups, i.e., qubits of a link do not get preferential treatment based on the time-slot they were created in.
- Each group of qubits to fuse must be entangled with qubits across distinct neighboring edges.
- A repeater maximizes the number of $n$-fusions that can be performed with the available links, followed by maximizing $(n-1)$-fusions, and so on.
- Qubits that are not part of any fusion group at a node are removed by performing Pauli-X basis measurements.
- The measurement results of each of the single-qubit Pauli-X measurements and the multi-qubit fusion measurements are communicated to Alice and Bob at the end of each time slot.

All fusion measurements and Pauli-X measurements at repeater nodes commute with one another. Hence they can be performed without coordination, yet synchronously, in each time slot, based on the protocol described above that uses local link-state information. Because all the qubits across the network in any $k$-time-slot block are measured (and hence destroyed), either with fusion or Pauli-X measurements, at the end of each block of $k$ time slots, Alice and Bob end up with potentially one or more shared GHZ states involving qubits that are held at Alice’s and Bob’s local memory registers. We define the entanglement generation rate as the number of GHZ states shared divided by $k$, expressed in the unit of GHZ states per time slot. Note that when $k = 1$, i.e., the $(n,1)$-GHZ protocol reduces to the $n$-GHZ protocol as presented in [9].

B. Mapping our protocol to a mixed-percolation problem

Let us consider our time-multiplexed $(n,k)$-GHZ protocol on a quantum network with an underlying topology given by a graph $G = (V,E)$ defined over a vertex set $V$ and edge set $E$. The link success probability per time slot $p$, and the success probability of each $n$-fusion $q$, translate respectively to effective bond and site occupation probabilities of a site-bond (mixed) percolation problem. The number of shared GHZ states after the measurements are performed at the end of each $k$-time-slot block equals the number of disjoint
Subgraphs shared between qubits held by Alice and Bob, in a graph whose nodes are individual qubits and whose edges are entangled links across network edges and edges drawn among qubits of fusion groups within repeater nodes as depicted in Fig. 1 [9]. In the appendix, we present a graphical construction, which shows that employing time multiplexing (i.e., \( k > 1 \)) simply changes the lattice on which this mixed-percolation problem is defined on. The percolation rules remain the same as those for the non-time multiplexed protocol described in [9].

In Fig. 2 we plot the critical boundaries for the site-bond percolation problem induced by the \( (4, k) \)-GHZ protocol with increasing time-multiplexing block length \( k \). The entanglement generation rate does not scale with distance in the super-critical region, i.e., above the critical boundary. The entanglement rate decays exponentially with the distance between Alice and Bob on the other side of the critical boundary. Increasing \( k \) expands the super-critical region, and hence affords distance-independent rates with progressively lower \( (p, q) \) values than is possible with the \( k = 1 \) protocol in [9]. Increasing \( k \) indefinitely would in principle indefinitely expand the critical boundary outwards. However, after \( k = 7 \) or so, for the square-grid topology, we notice diminishing returns of increasing \( k \) further. Further, the maximum \( k \) that can be supported would depend upon the maximum quantum memory coherence times. Recall that a coherence time of \( \frac{\tau_L}{\tau} + (k - 1) \) times slots (or this times \( \tau \) seconds) is necessary to support our protocol.

We observed that our described method to choose fusion groups at repeater nodes results in a larger super-critical region compared to a protocol that randomly groups links at a node without trying to maximize the number of \( n \)-fusions. However, we also believe that an improved version of our protocol is possible that time-orders link successes akin to [4].

C. Simulation results

All simulation results reported are for the \( (4, k) \)-GHZ protocol over the square-grid network, unless specified otherwise. We observed qualitatively similar trends for the \( (3, k) \)-GHZ protocol, but with admittedly smaller percolation regions than their \( (4, k) \) counterparts. In Fig. 3 we plot entanglement rate (in units of shared GHZ states between Alice and Bob, per time slot) as a function of link success probability \( p \) (fusion success probability \( q \)) for different values of fusion success probability (link success probability), both for the \( (3, k) \)- and \( (4, k) \)-GHZ protocols, for the cases of \( k = 2 \) and \( k = 8 \). As expected from the percolation regions shown in Section II-B when \( k \) goes from 2 to 8 the values of \( p \) and \( q \) where a sudden jump occurs in the rate decrease. This sharp transition in the rate at certain values of \( p \) and \( q \), and then saturating to a constant is an essence of percolation. In the flat region of the rate plots, the rate does not change with the distance between Alice and Bob. Below the percolation transitions, the rate diminishes exponentially with distance. The \( (p, q) \) pair at which the rate plots transition to a constant lie on the site-bond critical curve for that \( k \) value in Fig. 2. If we keep increasing \( (p, q) \) while keeping \( k \) constant, the rate always saturates at \( 1/k \) GHZ states per time slot. This is because increasing \( (p, q) \) beyond the transition point pushes the network deeper into the super-critical region, where there exists with high probability one unique giant connected (spanning) component.
shared between the qubits of Alice and Bob. As a result, in this region, at most one GHZ state is shared between the consumers at the end of each \( k \)-time-slot block, resulting in the rate of \( 1/k \) GHZ states per time slot. We observe that time multiplexing gets rid of the ‘turn around’ in the rate that was observed for the 3-GHZ protocol in [9]. The turn around occurs in the \((3,1)\)-GHZ protocol on the square grid network in [9] because if at least one of the two neighboring repeaters has four successful links, the repeater(s) with four links may not choose the link shared with the other repeater for fusion while the other repeater uses the shared link for fusion. This random selection rule ends up effectively disconnecting the two repeaters. However, for the \((3,k)\)-GHZ protocol, when \( k > 2 \), the repeaters can perform multiple fusions. As a result, the link which was left unused in the \((3,1)\)-GHZ protocol now gets used to in some other fusion with a high probability, keeping the network connected.

We keep \((p,q)\) constant and vary \( k \) in Fig. 4 to demonstrate that time-multiplexing helps when values of \( p \) and \( q \) are small. We observe that for a given \((p,q)\) pair, the rate increases with \( k \) until a certain value of \( k \), which we call \( k_{(p,q)} \), for that \((p,q)\) pair, at which point the rate hits \( 1/k_{(p,q)} \) GHZ states per time slot. This is the maximum rate allowed for by the underlying network topology for that \((p,q)\) pair with the \((n,k)\)-GHZ protocol. This behaviour can be explained using the percolation plots in Fig. 2. If we keep \((p,q)\) constant and start increasing \( k \), the network barely percolates at \((p,q)\) when \( k = k_{(p,q)} \), i.e., \((p,q)\) falls within the super-critical region for the first time when \( k = k_{(p,q)} \). If \( k \) is increased beyond \( k_{(p,q)} \), the rate starts decaying as \( 1/k \) since \((p,q)\) remains in the super-critical region for larger values of \( k \) and the number of shared GHZ states stays 1.

If a \((p,q)\) pair falls in between the critical boundaries for two consecutive \( k \) values \( k_1 \) and \( k_2 = k_1 + 1 \), then the optimal time-multiplexing block length \( k_{(p,q)} \) is \( k_2 \). For example, \((p,q) = (0.4,0.6)\) falls between the critical boundaries for \( k = 7 \) and \( k = 8 \) as seen in Fig. 2. This \((p,q)\) pair would give exponentially decaying rate with distance for \( k = 7 \) and distance independent rate for \( k \geq 8 \). The optimal value of time-multiplexing block length therefore, \( k_{(0.4,0.6)} = 8 \). If \( k \) is increased above 8, the rate will equal \( 1/k \), but will remain distance-independent. The optimum \( k_{(p,q)} \) depends on both the network topology and the \( n \) in the \((n,k)\)-GHZ protocol, the largest permissible fusion-group size.

Fig. 5(a) presents a heat-map of the entanglement rate optimized over the choice of \( k \). The blue region with rate very close to 1 GHZ state per time slot is the super-critical region for \( k = 1 \), the non-time multiplexed protocol. We did simulations only up to \( k = 8 \). This is why we observe non-zero rate only inside the super-critical region of \( k = 8 \), and hence the maroon region outside of the \( k = 8 \) super-critical region should be ignored. Fig. 5(b) shows the optimal time-multiplexing block length \( k_{(p,q)} \) for the entire \((p,q)\) space, again modulo the fact that our simulations stop at \( k = 8 \). Between the percolation curves for two consecutive \( k \) values \( k_1 \) and \( k_2 = k_1 + 1 \), \( k_{(p,q)} = k_2 \), as expected. This plot can be used by the network design engineer as a tool to decide the optimal degree of time-multiplexing if \( p \) and \( q \) are known.

In Fig. 6 we plot the entanglement generation rate for three different distances between the communicating parties, to show that the rate does not change with increasing (Manhattan) distance between Alice and Bob, since \((p,q) = (0.9,0.9)\) is within the super-critical regions of all values of \( k \), the time-multiplexing block lengths, between 1 and 8 that are taken.

### III. MEMORY DECOHERENCE

In Section II we evaluated the performance of our \((n,k)\)-GHZ protocol assuming ideal quantum memories that can hold qubits noiselessly for \( \tau_L + (k-1)\tau \) seconds. Our analysis showed that by increasing the time-multiplexing block length \( k \), we can in principle indefinitely expand the super-critical region of percolation ans hence achieve a distance-independent rate for arbitrarily low \( p \) and \( q \). But increasing \( k \) requires higher memory coherence times as said above. In this section, we consider more realistic quantum memories that have finite coherence times, and assess how the optimal time-multiplexing block length \( k_{(p,q)} \) is not only a function of \( p \) and \( q \), but also a function of the mean coherence time of a memory, i.e., how long a memory can (on an average) hold on to a qubit noiselessly.

#### A. The memory decoherence model

We assume a step-function heralded decoherence model wherein each qubit is held perfectly in a quantum memory for a (discrete) number of time slots that is an exponentially-distributed random variable with mean \( \mu \), and after that the qubit is measured off in the Pauli-X basis. This model is a good first-order approximation of common quantum memory technologies, such as color centers in diamond and trapped ions, which exhibit exponentially decaying qubit fidelities with a measurable time constant [14]. The protocol stays exactly as described in Section II-A wherein each repeater node only employs local link-state knowledge in forming its fusion groups, and performs X measurements on unused qubits within each \( k \) time slot block. The only difference between the protocol described in Section II-A is that there may be a few additional qubits within each \( k \)-time-slot block—the ones which happen to decohere—on which Pauli-X measurements are performed as well. But these decoherence-induced Pauli-X measurements are (a) performed before the fusion-groups are formed at the end of each \( k \) time-slot block, and (b) a repeater node does not have information about which qubits decohered at any neighboring node. Giving the repeater node this information would add a latency of \( \tau_L \) to the protocol, and more qubits could be lost to decoherence while waiting for this information to arrive, hence defeating the purpose of benefiting from time multiplexing.

Another thing to note is that our mention of ‘qubit’ in the paper should be interpreted as being logical (error-corrected) qubits, and not physical qubits. This is because loss on a logical qubit can be heralded with the error-detection property of a quantum code. This is an important assumption in
Fig. 3. Effect of time-multiplexing on entanglement generation rate (in units of shared GHZ states between Alice and Bob per time slot). (a) Rate vs. \( p \) for \((3, k)\)-GHZ protocol, (b) Rate vs. \( q \) for \((3, k)\)-GHZ protocol, (c) Rate vs. \( p \) for \((4, k)\)-GHZ protocol, (b) Rate vs. \( q \) for \((4, k)\)-GHZ protocol. For each of the four plots, we have chosen the cases of \( k = 2 \) and \( k = 8 \). Solid lines correspond to \( k = 2 \) and dashed lines are for \( k = 8 \). The rate always saturates to \( 1/k \) GHZ states per time slot. This is because, if the percolation condition is met (i.e., in the super-critical regime), there exists with high probability a unique single giant-connected (spanning) component, which translates to Alice and Bob sharing one entangled GHZ state among qubits held in their memory registers.

B. Results

In Fig. 7, we keep \( \mu \), the mean decoherence time of each qubit constant and vary the time-multiplexing block length \( k \) to plot the envelope of site-bond critical boundaries (solid black and dashed red lines) for our \((n, k)\)-GHZ protocol, assuming memory decoherence as described above. Note that, \( \mu/2 \) is the average time length for which a link survives after being created. We observe that decreasing \( \mu \) shrinks the envelope implying the network needs to have high \( p \) and \( q \) to achieve distance independent rate if memory coherence time is small. If \( k \) is increased for a fixed \( \mu \), the super-critical region expands until \( k \) reaches a threshold \( k^* (\mu) = \mu/2 \). When \( k > \mu/2 \), on an average, only the links created in the previous \( \mu/2 \) time-steps survive the memory decoherence. If a qubit of a link is lost at a repeater, since this information is heralded and known to the repeater, it uses it in forming its fusion groups. But this
information does not get communicated to the neighboring nodes. The neighboring node that holds the other qubit of the link whose one qubit is lost, it may unwittingly include that surviving qubit in one of its fusion groups, assuming the corresponding link is alive. This fusion thus does not contribute to expanding the reach of entanglement in the direction of the lost qubit. Hence, the remaining fusions in the network need to succeed with a higher probability \(<q\) compared to when \(<\mu=\mu/2\) to get at least one GHZ state shared between the consumers, shrinking the super-critical region.

We plot rate versus \(\mu\) for three fixed \(p,q\) and \(k=5\) in Fig 8a. The \((p,q) = (0.4,0.6)\) pair falls out of the super-critical regions of all \(\mu\) values when \(k = 5\), giving exponentially decaying rate. When \((p,q) = (1,1)\), the network always percolates. As a result, the rate is always \(1/k = 0.2\) GHZ states per time slot. For \((p,q) = (0.9,0.65)\), the rate monotonically increases with \(\mu\), and after \(\mu = 10\), the increase in rate shows diminishing returns. This \((p,q)\) pair is just outside or on the boundaries of the site-bond curves when \(\mu < 10\). It just enters the super-critical region when \(\mu = 10\), and stays inside, making the rate \(1/k = 0.2\) GHZ states per time slot for \(\mu > 10\).

Fig 8b shows rate as a function of time-multiplexing block length \(k\) for a fixed \(\mu = 10\). If \((p,q)\) is very deep into the super-critical region for all values of \(k\), such as when \((p,q) = (0.75,1)\), the rate isn’t affected by memory decoherence and follows the \(1/k\) curve. If \((p,q)\) is inside the super-critical region of only a few \(k\) values, the rate increases with \(k\) until \(k\) reaches a certain value, what we call \(k_{(p,q)}\). If we keep increasing \(k\) such that \(k > k_{(p,q)}\), the rate starts falling, eventually reaching 0. For example, when \((p,q) = (0.75,0.65)\), the rate is maximum for \(k=(0.75,0.65,10) = 4\).

IV. NETWORK DIVISION

In the \(n\)-GHZ protocol discussed in [9], which is the \((n,1)\)-GHZ special-case of the time-multiplexed protocol described in Section II-A, the Bell pairs shared between neighboring repeaters are fused using \(n\)-fusions at every repeater that is not a consumer. At the end of each time slot, the consumers end up with shared GHZ state(s) with some probability. The expected number of these shared GHZ states per time slot is the entanglement rate of the protocol. This protocol translates to a site-bond percolation problem on the lattice created after performing fusions such that the site and bond occupation probabilities of the percolation problem are the fusion and link success probabilities, i.e., \(q\) and \(p\) respectively. The number of disjoint components in the lattice that contain both consumers after performing the fusion gives the number of GHZ states shared between the consumers. In the supercritical regime of a most standard percolation problems, such as site, bond, and site-bond percolation, there is with high probability only one connected component whose size is proportional to the number of nodes in the network. This unique connected component is known as the giant connected component. This feature of percolation limits the rate of the protocol to one GHZ state per time slot in the super-critical regime. This upper-bound on the rate is independent of network topology. However, the maximum achievable rate for a given network is given by the min-cut of the network [10]. For example, with \(p = 1\), the square grid network should be able to achieve a rate of 4 GHZ states per slot. However, a randomized protocol as ours in this paper, would only achieve a rate of one GHZ state per slot, even for \(p = 1\). In this section, we show how spatial-division multiplexing can help go beyond 1 GHZ state per slot, even without time multiplexing (i.e., for \(k = 1\)), for non-trivial values of \(p\) and \(q\).

A. The network-division protocol

The rate of the \(n\)-GHZ protocol of [9] can be increased in the large \((p,q)\) region by enforcing more than one disjoint components to be created between Alice and Bob. In this section, we describe one possible method to achieve that. We divide the square-grid network into four regions as shown in Fig. 9. This method can also be seen as spatial-division multiplexing. We define region I (shown in green) such that it is a square region and Alice and Bob are located at its diagonally opposite corners. Region II (shown in red) is composed of three copies of region I. Region III is identical to region II in shape but on the opposite side of region I. The rest of the network (the white space in Fig. 9) constitutes region IV. The quantum repeaters on the boundaries of the four regions are not allowed to fuse links from different regions. This is equivalent to splitting every repeater on the region boundaries into two internal repeaters that act independent of one another, forming fusion groups within their own respective regions. For the square grid, each of these two internal repeaters can perform only a 2-fusion, i.e., a BSM. This rule effectively disconnects one region from its neighboring regions and each region yields up to one shared entangled states. If a \((p,q)\) value...
Fig. 5. (a) Entanglement rate $R_{\text{max}}$ achieved for every $(p, q)$ point maximized over the choice of $k$, for the $(4, k)$-GHZ protocol over a square-grid network; (b) The optimal time multiplexing block length $k_{(p,q)}$ plotted as a heat-map as a function of $p$ and $q$, which attains $R_{\text{max}}$ shown in (a) for every $(p, q)$ value. For both plots, our simulations stop at $k = 8$. Hence, $k_{(p,q)}$ does not exist in the white region of (b).

Fig. 6. Entanglement rate is plotted as a function of the time-multiplexing block length $k$ for $(p, q) = (0.9, 0.9)$, plotted for three different distance between the communicating parties Alice and Bob. Since $(p, q) = (0.9, 0.9)$ falls in the super-critical regions for the percolation problems for all $k \in \{1, \ldots, 8\}$, the rate is seen not to be dependent on the distance between the end users Alice and Bob.

Fig. 7. Site-bond percolation envelopes shown as black solid and red dashed lines when the quantum memories have finite coherence time, distributed exponentially with mean $\mu = 10$ and $\mu = 12$, respectively. Increasing memory coherence time improves the super-critical region, as expected. The site-bond percolation curves for different values of $k$ when $\mu = 10$ are shown as grey lines. The super-critical region expands with increasing $k$ till $k = \mu/2$ and starts shrinking if $k$ is increased further.

Note that, as Alice and Bob both can have at most one link per region each, the entangled state created in each region can only be a two-qubit state or a Bell state. This restriction in effect shrinks the percolation boundaries for each region (compared to that of the overall network), but for a high-enough $(p, q)$ opens up the possibility of multiple percolating regions and hence a rate that is higher than 1 GHZ state per slot.

This network division rule can be easily generalized to any regular network topology such that every region of the network has exactly one link each from Alice and Bob. The maximum rate achieved for this rule is equal to the node degree of the network topology, which is also equal to the min-cut of the

resulting more than one region to be in its respective critical region, the network would attain a distance independent rate that is at least the number of such (percolated) regions. The original $n$-GHZ protocol of [9] is performed inside each region separately. If $(p, q)$ is inside the super-critical region of region I of the protocol, i.e., the underlying lattice percolates from Alice to Bob, we can say that in region II, the lattice percolates from Alice to repeater C, then from C to D, and from D back to Bob. We can similarly divide the white region in smaller regions of the dimensions of region I and ensure percolation between Alice and Bob. We can conjecture that this construction can be extended to arbitrary locations of A and B.
network. If an $n$-GHZ protocol is performed on a regular network topology with node degree $d$ such that $n < d$, we would modify the fusion rule such that the neighbors of the consumers always use the links shared with the consumers, if successful, for fusion. If the neighbors do not end up using the links with the consumers, the entanglement rate immediately drops to zero in that region. In case of random graphs such as the configuration graph, although it is always possible to divide the network into different regions, the construction of these regions is not obvious. Even for the square-grid lattice, it is very likely that our region subdivisions are not optimal. Moreover, the optimal division of the network into regions may be a function of $p$ and $q$, and will also depend upon $k$, when we employ this space-division strategy to the time-multiplexed $(n, 1)$-GHZ protocol.

In recent work, Manna and Ziff analyzed a new percolation problem [18], wherein $m$ points separated by a distance proportional to the lattice size $L$ simultaneously connect together, or a single point at the center of a lattice connects to the boundary through adjacent connected points of a single cluster. We believe that a variant of such a percolation problem can help us study the percolation problem induced by our space-division-multiplexed protocol within each disconnected region.

### B. Simulation results of entanglement rates

In Fig. 10(a), we plot the entanglement rate as a function of $p$ for fixed $q$ values, and in Fig. 10(b), we plot the entanglement rate as a function of $q$ for fixed $p$ values. We observe that the network division protocol achieves the rate of 4 GHZ states per slot at $p = q = 1$, the maximum rate achievable in a square grid network. The rate plots show a percolation transition, but of a different critical exponent compared to the percolation seen earlier for the $(n, k)$-GHZ protocol. Each region has its own critical boundary in the $(p, q)$ space, which we have not evaluated. The total rate is sum of rates of the four regions. When the rate is above 2, e.g., for $(p, q) = (0.85, 9)$, it implies that at least two regions are in their respective super-critical regions.

Furthermore, we observe that for $(p, q)$ values just inside the super-critical region, the rate attained by dividing the network is lower than the undivided network. Hence, for low $(p, q)$ values, spatial division multiplexing is not always advisable. This is because that $(p, q)$ which was inside the super-critical region of the undivided network may fall out of the super-critical regions of all the four regions. On the
other hand, if \((p, q)\) is very high, spatial division multiplexing can afford far superior rates than the \(n\)-GHZ protocol. Note that time-multiplexed \(n\)-GHZ protocol can be run in each region to improve the super-critical site-bond region further. The exploration of this is being left for future work.

V. CONCLUSIONS AND DISCUSSION

It was recently shown that by allowing quantum switches—generalization of a quantum repeater that perform dynamically-switchable quantum measurements—in a quantum network to perform multi-qubit joint projective measurements in the GHZ basis, the quantum network can achieve distance-independent entanglement rate that can vastly surpass the exponentially decaying rate in a certain region of link success probability \(p\) and fusion success probability \(q\), called the super-critical region of a new mixed-percolation problem \([9]\). This parameter region is determined by the percolation theory. In this paper, we showed that this super-critical region can be expanded by increasing the effective \((p, q)\) by use of time multiplexing \([4]\). As a result, the time-multiplexed protocol attains distance-independent entanglement rate for values of \((p, q)\) for which the network would have experienced exponentially decaying rate if not for the utilization of time multiplexing. We define an \((n, k)\)-GHZ protocol, wherein a switch can perform a fusion measurement on up to \(n\) locally-held qubits, and it chooses which groups of qubits to fuse after every subsequent block of \(k\) time slots. For a given network topology, every \((p, q)\) pair has an optimal value of time multiplexing block length \(k_{(p,q)}\) associated with it. Increasing the time-multiplexing block length \(k\) beyond that only hurts the rate, as it starts decaying as \(1/k\) GHZ states per slot. If the \((p, q)\) parameters are known, we have designed a method to calculate the degree of time-multiplexing that should be used in the network to attain the maximum possible rate for our \((n, k)\)-GHZ protocol.

The rate achieved by our \((n, k)\)-GHZ based time-multiplexing protocol could be improved further by restricting the repeaters to mix-and-match links generated in specific groups of time-steps. For example, if \(k = 8\), then the repeaters could fuse links created in the first four time slots with each other and those created in time slots 5-8 with one another. This protocol would achieve the rate that is twice that of the \((n, 4)\)-GHZ protocol rate. But it also comes with a trade off: if \((p, q)\) is outside the super-critical region for \(k = 4\), dividing the links into two groups may not be the best idea. Ideally, one could do an integer partitioning of \(k\) such that there are \(n_k\) partitions and if the network percolates at \((p, q)\) for the smallest partition, the rate achieved by the partitioned protocol is \(n_k\) times the \((n, k)\)-GHZ protocol.

While time-multiplexing helps to improve the rate in low the \((p, q)\) regime, it doesn’t help much when \((p, q)\) is high. We observed that, for the \(n\)-GHZ protocol, i.e., the \(k = 1\) special case of our \((n, k)\)-GHZ protocol, space-division multiplexing, i.e., restricting the flow of entanglement to certain sub-regions of the network, boosts the rate for when there is too much connectivity in the network caused due to high \((p, q)\) values. It results in multiple independent entangled states to be created between Alice and Bob. We show that by strategically dividing the network into disconnected regions, it is possible to achieve the entanglement capacity (the min-cut rate) \([10]\)—the maximum achievable capacity of the network assuming ideal repeaters—at \(p = q = 1\). We believe that the min-cut rate of 4 GHZ states per slot for the square grid may be achievable for \((p, q)\) strictly less than \((1,1)\) with the use of optimal time and spatial multiplexing. This will have to involve the study of percolation properties of the network sub-regions, which could potentially benefit from some recent work in percolation theory \([18]\). We leave the exploration of this to future work.

Finally, we break the assumption of having ideal quantum memories with infinite coherence time and use realistic
quantum memories whose coherence times are exponentially distributed random variables with mean $\mu$. The super-critical region expands with increasing $k$ until it reaches the average link survival time, $\mu/2$. The region shrinks when $k > \mu/2$ as the network requires the fusions to succeed with a higher probability to produce end-to-end shared entangled states. We find that if the network is deep into the super-critical region of percolation for a given $k$, i.e., inside all the percolation curves for different time-multiplexing block lengths and a fixed $\mu$ memory, decoherence doesn’t hurt the rate. The rate decays as $1/k$, similar to the case when ideal memories are used. However, for other $(p, q)$ values, there is an optimal time-multiplexing block length $(k_{(p,q,\mu)})$, which is a function of $\mu, n$ and the network topology that achieves the maximum rate. The rate starts decaying if $k$ is increased beyond $k_{(p,q,\mu)}$.

Although our memory decoherence model is an improvement over using ideal memories, it is still a very simplified model. A better decoherence model would include links established with sub-unity fidelity at the outset, and realistic continuous-time noise mechanisms such as depolarization noise acting on each qubit while it is held in a quantum memory. Multi-path entanglement routing using BSMS on links with sub-unity fidelity was recently studied in [7] by optimizing the scheduling of BSMS and Bell state purification in the network. Studying the performance of the $(n, k)$-GHZ protocol under this decoherence model would require scheduling GHZ-state purifications and $n$-fusions in the network. Refs. [19], [20] study GHZ purification circuits, but they require quantum states of the same sizes to undergo purification. It may not always be possible to get two or more GHZ states of the same size shared between two repeaters. Hence, one needs to design GHZ purification circuits that can work with GHZ states of different sizes. One can go a step further and include gate imperfections, i.e., noisy gates to perform the GHZ entanglement using local operations and classical communications (LOCC), in the analysis and performance evaluation of entanglement distribution in a quantum network.

ACKNOWLEDGMENT

The authors thank Gayane Vardoyan for insightful discussions.

REFERENCES

[1] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, “Fundamental limits of repeaterless quantum communications,” Nat. Commun., vol. 8, p. 15043, 2017.
[2] M. Takeoka, S. Guha, and M. M. Wilde, “Fundamental rate-loss tradeoff for optical quantum key distribution,” Nat. Commun., vol. 5, p. 5235, Oct. 2014.
[3] M. Pant, H. Krovi, D. Towsley, L. Tassiulas, L. Jiang, P. Basu, D. Englund, and S. Guha, “Routing entanglement in the quantum internet,” npj Quantum Information, vol. 5, no. 1, pp. 1–9, 2019.
[4] P. Dhara, A. Patil, H. Krovi, and S. Guha, “Sub-exponential rate versus distance with time multiplexed quantum repeaters,” arXiv preprint arXiv:2105.01002, 2021.
[5] Y. Lee, E. Bersin, A. Dahlberg, S. Wehner, and D. Englund, “A quantum router architecture for high-fidelity entanglement flows in multi-user quantum networks,” arXiv preprint arXiv:2005.01852, vol. 3, 2020.
[6] B. Li, T. Coopsmans, and D. Elkouss, “Efficient optimization of cut-offs in quantum repeater chains,” in 2020 IEEE International Conference on Quantum Computing and Engineering (QCE). IEEE, 2020, pp. 158–168.
[7] M. Victora, S. Krastanov, A. S. de la Cerda, S. Willis, and P. Narang, “Purification and entanglement routing on quantum networks,” arXiv preprint arXiv:2011.11644, 2020.
[8] E. Van Milligen, J. I. Jacobson, A. Patil, D. Towsley, and S. Guha, “Entanglement distribution protocol utilizing local link state knowledge with time multiplexing,” unpublished.
[9] A. Patil, M. Pant, D. Englund, D. Towsley, and S. Guha, “Entanglement generation in a quantum network at distance-independent rate,” arXiv preprint arXiv:2005.07247, 2020.
[10] S. Pirandola, “End-to-end capacities of a quantum communication network.” Communications Physics, vol. 2, no. 1, pp. 1–10, 2019.
[11] [Online]. Available: https://github.com/linkheros/FusionProtocol

APPENDIX

In this section, we describe the percolation problem corresponding to the $(n, k)$-GHZ protocol. To generate the lattice on which we perform percolation -

- for each repeater pair, create $k$ bonds, corresponding to generation of $k$ links i.i.d with probability $p$.
- We treat every fusion as a site in the lattice. Hence, total number of sites is $kN$ where $N$ is the number of repeaters in the network.
- Each repeater can have maximum $k$ sites corresponding to $k$ potential fusions.
- In the protocol, every repeater tries to maximize the number of $n$-fusions. Hence, for a repeater pair, if $k_s$ out of $k$ links succeed, we assign these links to first $k_s$ sites at both repeaters.
- Each repeater randomly chooses a link from all successful links to a neighbouring repeater. Hence, we randomly pair these $k_s$ sites from both the repeaters. It means, we randomly assign neighbourhood to the $k_s$ sites.
- Randomly generated neighborhood doesn’t matter for sites corresponding to the failed links as these links never participate in fusions.
• A new lattice with new neighbourhood is created for every iteration of the protocol due to the dynamic or random way, in which links are chosen to undergo fusions. Once the lattice is generated, we perform the usual site-bond percolation calculations corresponding to the non-time multiplexed $n$-GHZ protocol. These rules for lattice generation are applicable to both regular and random graph networks.