Improved analytical bounds on delivery times of long-distance entanglement

Coopmans, T.; Brand, S.O.; Elkouss, D.

Improved analytical bounds on delivery times of long-distance entanglement

Tim Coopmans,[1,*] Sebastiaan Brand,[2,†] and David Elkouss[1,‡]

[1] QuTech, Delft University of Technology, The Netherlands

[2] Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands.

The ability to distribute high-quality entanglement between remote parties is a necessary primitive for many quantum communication applications. A large range of schemes for realizing the long-distance delivery of remote entanglement has been proposed, both for bipartite and multipartite entanglement. For assessing the viability of these schemes, knowledge of the time at which entanglement is delivered is crucial. For example, if the communication task requires two entangled pairs of qubits and these pairs are generated at different times by the scheme, the earlier pair will need to wait and thus its quality will decrease while being stored in an (imperfect) memory. For the remote-entanglement delivery schemes which are closest to experimental reach, this time assessment is challenging, as they consist of nondeterministic components such as probabilistic entanglement swaps. For many such protocols even the average time at which entanglement can be distributed is not known exactly, in particular when they consist of feedback loops and forced restarts. In this work, we provide improved analytical bounds on the average and on the quantiles of the completion time of entanglement distribution protocols in the case that all network components have success probabilities lower bounded by a constant. A canonical example of such a protocol is a nested quantum repeater scheme which consists of heralded entanglement generation and entanglement swaps. For this scheme specifically, our results imply that a common approximation to the mean entanglement distribution time, the 3-over-2 formula, is in essence an upper bound to the real time. Our results rely on a novel connection with reliability theory.

I. INTRODUCTION

The Quantum Internet is a vision of a world-wide network of nodes with the capability to transmit and process quantum information [1, 2]. Such a network enables tasks that are impossible classically, among which unconditionally-secure communication [3, 4], secure delegated computing [5] and extending the baseline of telescopes [6]. A primitive for such tasks is entanglement between remote nodes. Several schemes have been proposed for generating long-distance entanglement, all making use of intermediate nodes called quantum repeaters [7]. These proposals include chains of quantum repeaters [7–9] and generalizations to two-dimensions for serving multiple users [10–14].

Knowledge of the time that quantum repeater schemes take to deliver entanglement is highly relevant, for several reasons. Most evidently, the delivery rate should be sufficiently high for the application. Secure communication over video, for example, requires transmission rates of at least hundreds of kbits per second [15]. Furthermore, for the repeater proposals which make use of quantum memories and do not rely on error correcting codes, i.e. the ones that are closest to experimental reach, the delivery time influences the quality of the produced entanglement. The reason for this is that in these schemes, an entangled pair that is generated often needs to wait for another pair before the scheme can continue, and decoheres in memory while waiting. In addition, some memory types suffer from effects which are effectively time-dependent, such as noise which is induced each time the quantum processor attempts to generate remote entanglement [16], while for others the probability of extracting the state degrades over time [17]. Thus, the quality of the produced entanglement is a function of the time its generation takes. This implies that knowledge of the delivery time is crucial for assessing the viability of schemes for long-distance entanglement distribution using near-term hardware.

Analysis of the delivery time is generally challenging for the entanglement-distribution schemes that are closest to experimental reach because they consist of probabilistic components. The completion time of a such a scheme is not a single number but instead a random variable, which for many schemes has a complex structure due to the feedback loops and restarts. Although numerically, progress has recently been made in determining the completion time for increasingly larger networks [14, 18–22], numerical approaches provide only limited intuition and moreover are demanding in computation time when performing large-scale optimization over many network designs and hardware parameters. For this reason, analytical results are more convenient.

Unfortunately, due to the complexity of the problem, even the average completion time is known exactly only in limited cases: for quantum repeater chains consisting of at most four repeater nodes [19, 23] and a star network with a single node in the center and an arbitrary number of leaves [10]. For larger networks, analytical results only include approximations or loose bounds on the mean entanglement delivery time [24]. The approximations are based on the assumption that the success probabilities of some of the network components are very small [25–28] or close to 1 [24, 29, 30]. Neither approximations are ideal, since some success probabilities can be boosted by techniques such as multiplexing, while others are bounded well below 1 for some setups[31]. Indeed, numerics have shown for some of the approximations that they become increasingly bad as the size of the network grows [19, 20]. Another scenario in which the completion time probability distribution is brought back to a known form includes the discarding of entanglement [32, 33]. See [34] for a review of the completion time analysis for entanglement distribution schemes.

* t.j.coopmans@tudelft.nl
† s.o.brand@liacs.leidenuniv.nl
‡ d.elkousscoronas@tudelft.nl

A canonical use case which has found particularly much application is a symmetric nested repeater scheme [7, 35] where at each nesting level two entangled pairs of qubits, spanning an equal number of nodes, are connected. Consequently, the entanglement span doubles at each nesting level. For this scheme, it was empirically known [36] that for small success probabilities of connecting the pairs, the average time to in-parallel create both required initial pairs at each nesting level is roughly 3/2 times the average time for a single pair. This results in an approximation to the average completion time of the repeater scheme which is known as the 3-over-2 formula and has been frequently used since [8, 9, 25, 29, 36–49]. Analytically finding the exact factor, for an arbitrary number of nesting levels and for any value of the success probabilities, has been an open problem for more than ten years [25].

In this work, we provide analytical bounds on the completion time which not only improve significantly upon existing bounds, but also show *how good* some of the previous approximations are because the bounds become exact in the small probability limit. To be precise, we give analytical bounds on the mean and quantiles of the completion time random variable for entanglement-distributing protocols which are constructed of probabilistic components whose success probability can be bounded by a constant from below. This includes feedback loops in which failure of one component requires restart of other components, as long as no two components wait for the same other component to finish. Regarding the symmetric nested repeater protocol, our bounds imply that the 3-over-2 approximation is, in essence, an upper bound to the mean completion time, rigorously rendering analyses based on this approximation pessimistic. Other protocols we can treat include nested repeater chains with distillation and multipartite-entanglement generation schemes [10, 14, 50], among others.

This work is organized as follows. First, in Sec. II we describe the class of protocols our bounds apply to and introduce concepts from reliability theory we will use in the bounds' derivation. Sec. III contains our main results: analytical bounds on the mean completion time of such protocols and the tail of its probability distribution. Next, we obtain improved bounds with respect to existing work by applying these results to two use cases: a nested quantum repeater chain (Sec. IV) and a quantum switch in a star network (Sec. V). We prove the main results in Sec. VI and finish with a discussion in Sec. VII.

II. PRELIMINARIES

A. Protocols

The protocols considered in this work aim to generate bipartite or multipartite entanglement between remote parties. We will refer to bipartite entanglement as a 'link'. We consider protocols that are constructed from two building blocks: GENERATE and RESTART-UNTIL-SUCCESS. We treat them individually.

First, by GENERATE we refer to heralded generation of fresh entanglement. For simplicity, we will assume that the entanglement is bipartite and we will refer to such entanglement as an 'elementary link'. In our model, entanglement generation is performed in discrete attempts of fixed duration, each of which succeeds with a given constant probability p_{gen} [8]. The success is heralded, i.e. the nodes are aware which attempts fail and which succeed. The duration of a single attempt equals L/c, where L is the distance between the nodes and c is the speed of light in the transmission medium. We use L/c as the unit of time. As a consequence, the completion time of entanglement generation is a discrete random variable following the geometric distribution:

$$\Pr(T_{\mathrm{gen}} = t) = \begin{cases} p_{\mathrm{gen}}(1-p_{\mathrm{gen}})^{t-1} & \text{if } t \geq 1 \text{ is an integer} \\ 0 & \text{otherwise.} \end{cases} \quad . \tag{1}$$

We will denote the mean of this distribution by $\mu_{\mathrm{gen}} = 1/p_{\mathrm{gen}}$.

We will also consider the exponential distribution, which is the continuous analogue of the geometric distribution and is defined as follows: if X follows the exponential distribution with parameter $\lambda > 0$, then

$$\Pr(X > x) = e^{-\lambda x} \tag{2}$$

for any real number $x \geq 0$. For small p_{gen}, the completion time of entanglement generation is sometimes approximated by an exponential random variable $T_{\mathrm{gen}}^{\mathrm{approx}}$ with the same mean, which is achieved by setting $\lambda = 1/\mu_{\mathrm{gen}}$.

Next, we explain RESTART-UNTIL-SUCCESS by describing two of its instantiations, regarding entanglement swapping and entanglement distillation.

By an entanglement swap at node M, we refer to the operation which converts two links, one between nodes A and M and one between M and B, into a single long-distance link between A and B. We model the entanglement swap success with probability $0 < p_{\mathrm{swap}} \leq 1$, which is a constant that is independent of the states upon which the swap acts. In case of failure, both input links are lost. We model fusion, the generalization of the entanglement swap which converts more than 2 input links to a multipartite entangled state, in similar fashion.

Entanglement distillation is the probabilistic conversion of two low-quality links shared between two nodes to a single high-quality link between the same two nodes [51, 52]. The success probability of distillation depends on the states of the two links, and is lower bounded by $\frac{1}{2}$ for the schemes considered here. Similarly to the case of entanglement swapping, the two input links are lost if the distillation step fails.

We assume that the durations of the entanglement swap, fusion, and distillation operations are negligible.

In general, we use the term RESTART-UNTIL-SUCCESS for an operation which takes entanglement as input, performs a probabilistic operation onto it, and demands the regeneration of the input entanglement in the case of failure. Its success probability can be a function of properties of the input entanglement, such as its quality or its delivery time, but it may also be a constant. By SWAP-UNTIL-SUCCESS and DISTILL-UNTIL-SUCCESS, we refer to instantiations of RESTART-UNTIL-SUCCESS where the probabilistic operation is entanglement swapping and entanglement distillation, respectively.

The protocols we consider in this work are composed from heralded entanglement generation and RESTART-UNTIL-SUCCESS as subprotocols, with the restriction that the distinct RESTART-UNTIL-SUCCESS protocols do not compete for the same resources. That is, no pair of subprotocols waits for the same link before proceeding. This corresponds to the protocols where the dependency graph of the inputs and outputs of the subprotocols is a tree. Fig. 1 shows examples such protocols.

B. Probability theory and the NBU property

In this work, we will make extensive use of a class of probability distributions called new-better-than-used (NBU), which have been studied in the context of reliability theory and life distributions [53]. In order to mathematically define new-better-than-used, we first revisit some notions from probability theory. All random variables in this work that are continuous have the positive reals as domain, i.e. a continuous random variable X with $\Pr(X < 0) = 0$. The cumulative distribution function (CDF) of random variable X is $x \mapsto \Pr(X \leq x)$, and the co-CDF is $x \mapsto \Pr(X > x)$. This co-CDF is also referred to as the survival function or the *reliability*, since it states the probability that X will survive at least up to time x. The residual life distribution of X is given by the conditional probability $\Pr(X > x + y | X > y)$ and describes the time that X will survive at least up another interval x given that it has already survived time y. We now say that a real-valued random variable X is new-better-than-used (NBU) or that it has the NBU property if its residual life distribution is upper bounded by the original reliability, i.e.

$$\forall x, y \geq 0: \qquad \Pr(X > x + y | X > y) \leq \Pr(X > x). \tag{3}$$

Intuitively, new-better-than-used random variables describe ageing over time. As an example, consider the lifetime of a car: the probability that an old car (one that is already y years old) will survive another x years is smaller than the probability that a brand new car will reach the age of x years.

For clarity, we separately state the definition of NBU, where we use an expression equivalent to eq. (3) for convenience of our proofs later on.

Definition 1. *A real-valued random variable X with $\Pr(X < 0) = 0$, is called new-better-than-used (NBU) if*

$$\forall x, y \geq 0: \qquad \Pr(X > x + y) \leq \Pr(X > x) \cdot \Pr(X > y).$$

It is called new-worse-than-used (NWU) if the reverse inequality holds.

We give two examples of NBU distributions.

Example 1. *A delta-peak distribution $\Pr(X = x_0) = 1$ for some fixed $x_0 \geq 0$ is NBU, since*

$$\Pr(X > x)\Pr(X > y) = \begin{cases} 1 & \textit{if } x < x_0 \text{ and } y < x_0 \\ 0 & \textit{otherwise} \end{cases}$$

while

$$\Pr(X > x + y) = \begin{cases} 1 & \textit{if } x + y < x_0 \\ 0 & \textit{otherwise.} \end{cases}$$

Since $x+y < x_0$ implies $x < x_0$ and $y < x_0$ for any $x, y \geq 0$, we see that $\Pr(X > x + y) \leq \Pr(X > x)\Pr(X > y)$ and thus X is NBU.

Example 2. *The exponential distribution, defined in eq. (2), satisfies $\Pr(X > x + y) = \Pr(X > x)\Pr(X > y)$ for all $x, y \geq 0$ and is therefore both NBU and NWU.*

Lastly, we will use the notion of stochastic dominance.

Definition 2. *Let X and Y be two random variables with domains D_X and D_Y, both subsets of the real numbers. We say that X stochastically dominates Y and write $X \geq_{\text{st}} Y$ if*

$$\Pr(X > z) \geq \Pr(Y > z)$$

for all $z \in D_X \cap D_Y$.

In particular, we will use the following lemma, which states that stochastic dominance of one random variable over the other implies an ordering of their means.

Lemma 1. *Let X and Y be two random variables with domain $[0, \infty)$. If $X \geq_{\text{st}} Y$, then $E[X] \geq E[Y]$.*

Proof. The lemma directly follows from the definition of stochastic dominance, together with the fact that the mean of X can be written as an integral over the co-CDF,

$$E[X] = \int_0^\infty \Pr(X > x) dx,$$

and similarly for Y. □

III. MAIN RESULTS

In this section, we give our main results in Prop. 1 and 2: bounds on the completion time distribution for protocols composed of elementary-link generation (GENERATE) and RESTART-UNTIL-SUCCESS operations. The proofs to the main results can be found in Sec. VI.

Our results bound continuous completion times, whereas the completion time of elementary-link generation is the discrete random variable T_{gen} (see Sec. II). Therefore, before starting our main result we first remark that T_{gen} is stochastically dominated by a continuous NBU random variable we denote as $T_{\text{gen}}^{\text{upper}}$.

Lemma 2. *The completion time T_{gen} of elementary-link generation is stochastically dominated (Def. 2) by the continuous random variable $T_{\text{gen}}^{\text{upper}} = 1 + T_{\exp}$ where $T_{\exp}$ is exponentially distributed with parameter $\frac{-1}{\log(1-p_{\text{gen}})}$. That is,*

$$\begin{aligned}\Pr(T_{\text{gen}} > t) &\leq \Pr\big(T_{\text{gen}}^{\text{upper}} > t\big) \\ &= \begin{cases} 1 & \text{if } 0 \leq t \leq 1 \\ \exp\left((t-1)/\log(1-p_{\text{gen}})\right) & \text{if } t \geq 1 \end{cases}\end{aligned}$$

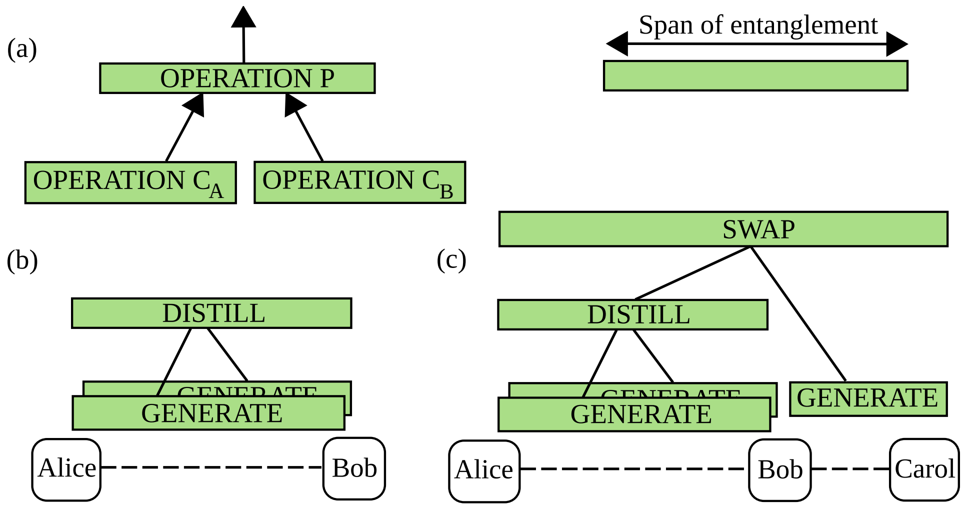


FIG. 1: The results in this work bound the completion time of any entanglement-distribution protocol which can be visualized as a tree. **(a)** In such a tree, each vertex is labelled by the operation (P) that should be performed as soon as the operations on the vertex's children (C_A and C_B) have finished. In case the operation fails, both children start regenerating entanglement, possibly by recursively having their children regenerate entanglement. This procedure is repeated until the operation P succeeds. **(b)** Example protocol on two nodes, Alice and Bob, which consists of performing heralded entanglement generation (GENERATE) twice in parallel, followed by entanglement distillation (DISTILL) on the two freshly generated links. In case of failure of the distillation attempt, both links are lost, in which case the protocol restarts. This procedure is repeated until the distillation attempt succeeds. **(c)** Example protocol on three nodes. Alice and Bob perform the protocol from (b), and in parallel Bob and Carol perform heralded entanglement generation. As soon as both have finished, Bob performs an entanglement swap (SWAP). This procedure is repeated until the swap succeeds.

The mean of T_{gen} is upper bounded by the mean of $T_{\mathrm{gen}}^{\mathrm{upper}}$ which is given by

$$\mu_{\mathrm{gen}}^{\mathrm{upper}} = 1 - \frac{1}{\log(1-p_{\mathrm{gen}})} = \frac{1}{p_{\mathrm{gen}}} + \frac{1}{2} + O(p_{\mathrm{gen}}) \quad (4)$$

where $O(p_{\mathrm{gen}})$ contains terms that scale with p_{gen} or powers of it. The means of T_{gen} and $T_{\mathrm{gen}}^{\mathrm{upper}}$ differ only slightly, both in difference and in ratio:

$$0 \leq \mu_{\mathrm{gen}}^{\mathrm{upper}} - \mu_{\mathrm{gen}} \leq \frac{1}{2} \text{ and } 1 \leq \frac{\mu_{\mathrm{gen}}^{\mathrm{upper}}}{\mu_{\mathrm{gen}}} \leq 1 + \frac{p_{\mathrm{gen}}}{2} \quad (5)$$

for any $p_{\mathrm{gen}} \in [0,1]$. Moreover, $T_{\mathrm{gen}}^{\mathrm{upper}}$ is NBU.

As consequence of Lemma 2, we may assume that the duration of elementary-link generation is described by $T_{\mathrm{gen}}^{\mathrm{upper}}$ if we are looking for upper bounds on a protocol's completion time. Indeed, an upper bound on the co-CDF or the mean of the resulting completion time will automatically also become an upper bound on the real completion time (see Def. 2 and Lemma 1).

Now let us state our bounds on continuous completion times. For legibility, we first state a special case of our main result: the scenario where a SWAP-UNTIL-SUCCESS operation with constant success probability is performed on two quantum states. We assume that the time it takes until a state is produced is a random variable, and that this random variable is the same for both input states; that is, their completion times are independent and identically distributed.

Completion time of swapping: two states & IID

Proposition 1. *Consider the time T_{output} of a* SWAP-UNTIL-SUCCESS *protocol with constant success probability p, acting on two quantum states, produced with identically-distributed independent completion times T_{input}. If T_{input} is a continuous random variable and it is NBU (Def. 1), then:*

(a) T_{output} is NBU;

(b) the mean of T_{output} is upper bounded as

$$E[T_{\mathrm{output}}] \leq \frac{3E[T_{\mathrm{input}}]}{2p};$$

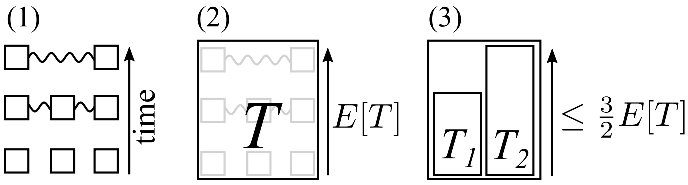
(c) for all t, the probability that T_{output} takes longer than t timesteps decays exponentially fast:

$$\Pr(T_{\mathrm{output}} > t) \leq \exp\left(p - \frac{2pt}{3E[T_{\mathrm{input}}]}\right)$$

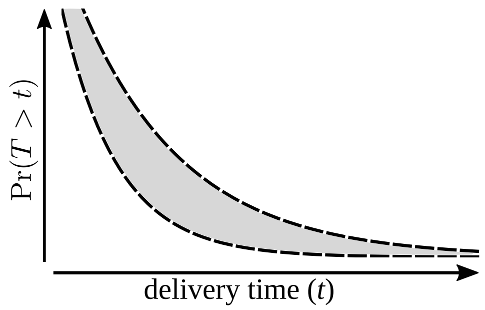
while it is lower bounded as

$$\Pr(T_{\mathrm{output}} > t) \geq \exp\left(\frac{-2pt}{3E[T_{\mathrm{input}}]} \cdot \frac{1}{1-p}\right).$$

(d) in the limit $p \to 0$, the normalized completion time $T_{\mathrm{output}}/E[T_{\mathrm{output}}]$ approaches the exponential distribution with mean 1, and thus $E[T_{\mathrm{output}}] \cdot 2p/(3E[T_{\mathrm{input}}]) \to 1$.



(a) Consider an entanglement distribution process (1), whose completion time is a random variable T and has mean $E[T]$ (2). If T is NBU, completing two such independent and identically distributed processes in parallel has a mean time $\leq \frac{3}{2} \cdot E[T]$ (3).



(b) The probability distribution of the delivery time of entanglement distribution processes can be bounded by exponentially-fast decaying lower and upper bounds.

FIG. 2: Visual overview of this work's bounds on the completion time of entanglement distribution protocols. The first result (2a) is a bound on the mean completion time of two parallel entanglement distribution processes, given that these processes possess the NBU property (Def. 1). Our second result (2b) is a two-sided bound on the probability distribution of the completion time of such processes.

The bounds from Prop. 1 are visually depicted in Fig. 2.

Although Prop. 1 regards a SWAP-UNTIL-SUCCESS protocol, it also finds application to DISTILL-UNTIL-SUCCESS, which has nonconstant success probability:

Remark 1. *Consider Prop. 1 where* SWAP-UNTIL-SUCCESS *is replaced by* DISTILL-UNTIL-SUCCESS. *Note:*

(a) Prop 1(a)-(c) still hold in case the quantum states produced with completion times T_{input} do not decohere over time, because then the distillation success probability p is a constant, independent of the production times of the input states;

The success probability of distillation is general lower bounded by $1/2$, resulting in

(b) $E[T_{\text{output}}] \leq 3E[T_{\text{input}}]$.

Since the upper bound in Prop 1(c) is monotonically decreasing in p in the regime $t \geq 3E[T_{\text{input}}]/2$, we may replace p by its lower bound $1/2$ to obtain:

(c) for $t \geq 3E[T_{\text{input}}]/2$, we have

$$\Pr(T_{\text{output}} > t) \leq \exp\left(\frac{1}{2} - \frac{t}{3E[T_{\text{input}}]}\right).$$

Prop. 1 is a special case of a more general version of Prop. 2 for RESTART-UNTIL-SUCCESS protocols that act on two or more quantum states whose completion times are independent, but not necessarily identically distributed.

General case: completion time of RESTART-UNTIL-SUCCESS protocol

Proposition 2. *Consider the time T_{output} of a* RESTART-UNTIL-SUCCESS *protocol with constant success probability p, acting on $n \geq 2$ quantum states, produced with independent completion times $T_1, \ldots, T_n$, which need not be identically distributed. Suppose that each of T_{output} and $T_1, \ldots, T_n$ is a continuous random variable. Denote $m = E[\max(T_1, \ldots, T_n)]$. If all $T_1, \ldots, T_n$ are NBU (Def. 1), then:*

(a) T_{output} is NBU;

(b) the mean of T_{output} equals $E[T_{\text{output}}] = m/p$;

(c) for all t, the probability that T_{output} takes longer than t timesteps is exponentially bounded from above as

$$\Pr(T_{\text{output}} > t) \leq \exp\left(p - \frac{p \cdot t}{m}\right).$$

while it is bounded from below by

$$\Pr(T_{\text{output}} > t) \geq \exp\left(\frac{-p \cdot t}{m} \cdot \frac{1}{1-p}\right).$$

(d) in the limit $p \to 0$, the normalized completion time $T_{\text{output}}/E\left[T_{\text{output}}\right]$ approaches the exponential distribution with mean 1, and thus $E[T_{\text{output}}] \cdot p/m \to 1$.

(e) We have

$$\max_{1 \leq j \leq n} E[T_j] \leq m \leq \sum_{j=1}^{n} E[T_j].$$

(f) In case all T_j are identically distributed with mean $E[T]$, then a tighter bound than (e) exists:

$$1 \leq \frac{m}{E[T]} \leq n - 1 + \frac{1}{n}.$$

We finish this section by generalizing Remark 1.

Remark 2. *Consider a* RESTART-UNTIL-SUCCESS *protocol whose success probability is lower bounded by a constant c. Then the upper bounds in Prop. 2(e) and (f) still hold, while Prop. 2(b) and (c) can respectively be replaced by $E[T_{\text{output}}] \leq m/c$ and $\Pr(T_{\text{output}} > t) \leq \exp\left(c - \frac{ct}{m}\right)$ for $t \geq m$.*

In the next sections, we give two use cases for the bounds derived in this section: a quantum repeater chain scheme and a quantum switch protocol.

IV. FIRST APPLICATION: NESTED QUANTUM REPEATER CHAIN

In this section, we apply our bounds on the completion time of entanglement distribution protocols to an extensively-studied nested repeater chain protocol [7, 35]. We explain the protocol for the case where the number of segments is 2^n for some integer $n \geq 0$ (i.e. the chain consists of $2^n + 1$ nodes). See also Fig. 3. If $n = 0$, then the network consists of two end nodes only (no repeaters), which use heralded entanglement generation (see Sec. II) to generate a single elementary link. If $n > 0$, then the chain has a middle node (since the number of segments is even). In parallel, a 2^{n-1}-hop-spanning link is produced on the left side of the middle node, as well as a link on its right side. As soon as both links have been prepared, the middle node performs an entanglement swap to convert the two links into a single 2^n-hop-spanning link. This scheme can also be extended with one or multiple rounds of entanglement distillation at each nesting level, in a nested fashion [7].

The exact completion time distribution of the nested repeater scheme has so far not been analytically found beyond the single-repeater case. The problem was first fully explained by Sangouard et al. [25], although it was already partially described in earlier work [36–38]. Sangouard et al., remarked that while the completion time of elementary-link generation at the bottom level follows a well-known distribution (the geometric distribution, Sec. II), this is no longer the case for higher levels.

To circumvent this issue, many have resorted to approximating the probability distribution at each level with an exponential distribution, combined with the small-probability assumptions $p_{\text{swap}} \ll 1$ and $p_{\text{gen}} \ll 1$. This approximation leads to an expression for the mean entanglement delivery time as follows. At each nesting level, the protocol can only continue if both input states to the entanglement swap have been produced. Mathematically, this is expressed as the maximum of the delivery time of the two links. The mean of the maximum of two independent and identically distributed (i.i.d.) exponential random variables with mean μ is $\frac{3}{2} \cdot \mu$. Next, if the swap success probability is p_{swap}, then on average $1/p_{\text{swap}}$ attempts are needed until success. Thus, for each nesting level, the mean entanglement delivery time should be multiplied by a factor $3/(2p_{\text{swap}})$, resulting into an expression for the mean delivery time known as the *3-over-2-approximation*:

$$\left(\frac{3}{2p_{\text{swap}}}\right)^n \cdot \frac{1}{p_{\text{gen}}}. \tag{6}$$

The 3-over-2 approximation was first used by Jiang et al.[36], who mentioned that the factor 3/2 agreed well with simulations in the small-probability regime. Since then, the approximation has been frequently used [8, 9, 25, 29, 37–49].

However, the quality of this approximation is not known exactly and has only been only very loosely bounded, as follows. As noted by Sangouard et al. [25], the mean of the maximum of two nonnegative i.i.d random variables with mean μ is lower bounded by μ and upper bounded by 2μ. These bounds correspond to the scenario where one waits only for a single link to be ready, or for both links to be prepared sequentially, respectively. Consequently,

$$\left(\frac{1}{p_{\text{swap}}}\right)^n \cdot \frac{1}{p_{\text{gen}}} \leq E[T] \leq \left(\frac{2}{p_{\text{swap}}}\right)^n \cdot \frac{1}{p_{\text{gen}}}. \tag{7}$$

Now we use Markov's inequality, $\Pr(T \geq t) \leq E[T]/t$, which can be rephrased

$$\Pr(T > t) \leq E[T] \cdot \frac{1}{t+1}, \tag{8}$$

since T only takes integral values. Substituting $E[T]$ by its upper bound from eq. (7) leads to

$$\Pr(T > t) \leq \left(\frac{2}{p_{\text{swap}}}\right)^n \cdot \frac{1}{p_{\text{gen}}} \cdot \frac{1}{t+1}. \tag{9}$$

Both the mean bound from eq. (7) and the tail bound from eq. (9) are quite loose bounds, see Fig. 4 and 5. Only recently, it was shown analytically by Kuzmin and Vasilyev that the factor 3/2 from eq. (6) is exact in the limit of vanishing swap success probability, and moreover that the delivery time probability distribution after an entanglement swap in this limit is indeed an exponential distribution [26].

Our bounds from Sec. III allow us to go beyond these results. In particular, we show the following. First, we analytically show that the 3-over-2 approximation is, in essence, an *upper bound* to the mean completion time. This implies that the 3-over-2 approximation is pessimistic, confirming numerical simulations [19, 29]. Next, we derive two-sided bounds on the tail of the probability distribution of the repeater chain's completion time. Both the mean bound and the tail bounds coincide in the limit of vanishing success probabilities. We give the bounds below and plot them in Fig. 4 (mean bounds) and Fig. 5 (tail bounds).

Proposition 3. *Consider the completion time T_n of an equally-spaced, symmetric nested repeater scheme (no distillation) on 2^n segments, such as the example in Fig. 3 for $n = 2$. If $n > 0$, then:*

(a) the mean completion time is upper bounded as

$$E[T_n] \leq \left(\frac{3}{2p_{\text{swap}}}\right)^n \cdot \mu_0.$$

Here, μ_0 is the mean of any real-valued NBU random variable which stochastically dominates the completion time T_{gen} of elementary-link generation. In case the elementary-link generation is modelled as discrete attempts which succeed with probability p_{gen}, then we choose $T_{\text{gen}}^{\text{upper}}$ for this random variable (see Lemma 2), resulting in

$$\mu_0 = E[T_{\text{gen}}^{\text{upper}}] = 1 - \frac{1}{\log(1 - p_{\text{gen}})}.$$

If instead the completion time of elementary-link generation is described by the exponentially-distributed random variable $T_{\text{gen}}^{\text{approx}}$ (see Sec. II A), which is NBU itself, then $\mu_0 = E[T_{\text{gen}}^{\text{approx}}] = 1/p_{\text{gen}}$. By Lemma 2, the two models' means only differ slightly: $0 \leq E[T_{\text{gen}}^{\text{upper}}] - E[T_{\text{gen}}^{\text{approx}}] \leq \frac{1}{2}$ and $1 \leq E[T_{\text{gen}}^{\text{upper}}]/E[T_{\text{gen}}^{\text{approx}}] \leq 1 + p_{\text{gen}}/2$.

(b) the mean completion time is lower bounded as

$$E[T_n] \geq \frac{1}{p_{\rm swap}} \cdot \left(\frac{3-2p_{\rm swap}}{p_{\rm swap}(2-p_{\rm swap})}\right)^{n-1} \cdot \nu_0.$$

Here, ν_0 is the mean time until the latest of two parallel elementary-link generation processes has finished. In case elementary-link generation is modelled as discrete attempts which succeed with probability $p_{\rm gen}$, then

$$\nu_0 = \frac{3-2p_{\rm gen}}{p_{\rm gen}(2-p_{\rm gen})}$$

while if its completion time is modelled by an exponential distribution, then $\nu_0 = 3/(2p_{\rm gen})$.

(c) the co-CDF of T_n differs from the co-CDF of an exponential distribution by at most a factor $\exp(p_{\rm swap})$ *from above,*

$$\Pr(T_n > t) \leq \exp(p_{\rm swap}) \cdot \exp\left(-\frac{p_{\rm swap} \cdot t}{m_{\rm upper}}\right)$$

while it is lower bounded as

$$\Pr(T_n > t) \geq \exp\left(\frac{-p_{\rm swap} \cdot t}{m_{\rm lower}} \cdot \frac{1}{1-p_{\rm swap}}\right).$$

Here, we have denoted

$$m_{\rm upper} = \frac{3}{2} \cdot \left(\frac{3}{2p_{\rm swap}}\right)^{n-1} \cdot \mu_0$$

and

$$m_{\rm lower} = \left(\frac{3-2p_{\rm swap}}{p_{\rm swap}(2-p_{\rm swap})}\right)^{n-1} \cdot \nu_0$$

where μ_0 and ν_0 are given in Prop. 3(a) and (b).

(d) in the limit where both $p_{\rm swap} \to 0$ and $p_{\rm gen} \to 0$, the normalized random variable $T_n/E[T_n]$ follows the exponential distribution with mean 1, and moreover

$$\lim_{p_{\rm swap}\to 0, p_{\rm gen}\to 0} E[T_n]/L_n = 1$$

with

$$L_n = \left(\frac{3}{2p_{\rm swap}}\right)^n \cdot \frac{1}{p_{\rm gen}}.$$

(e) If the completion time of elementary-link generation is described by the exponentially-distributed $T_{\rm gen}^{\rm approx}$, then T_n is NBU, while if it is modelled as discrete attempts, then T_n is stochastically dominated (Def. 2) by an NBU random variable which satisfies the bounds in items (a-c).

Most statements in Prop. 3 directly follow by applying Prop. 1 in Sec. III iteratively over the number of nesting levels. In particular, a useful feature following from Prop. 1(a) is that at each nesting level, the completion time possesses the NBU property (Def. 1). Consequently, the mean upper bound in Prop. 1(c), which is only applicable to NBU random variables, can be used at each nesting level. Only the lower bound in (b) and the expression for $m_{\rm lower}$ in (c) do not follow from Prop. 1. These can be found by noting that the maximum of two sums dominates a single sum whose length is the maximum of the original two sum lengths. We give the full proof in Appendix B.

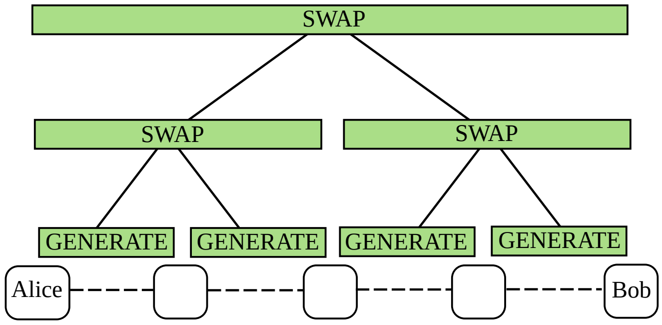


FIG. 3: Schematic of a nested repeater protocol on five nodes ($n = 2$ nesting levels) The figure depicts the protocol for delivering entanglement between remote parties Alice and Bob through three repeater nodes. At the start of the protocol, all nodes attempt to generate an elementary link with each of their neighbors in parallel. An entanglement swap is performed once the two leftmost links are ready, and similarly for the two rightmost links. Once both swaps have succeeded (failure requires regeneration of the involved links), the middle node performs an entanglement swap, which yields entanglement between Alice and Bob.

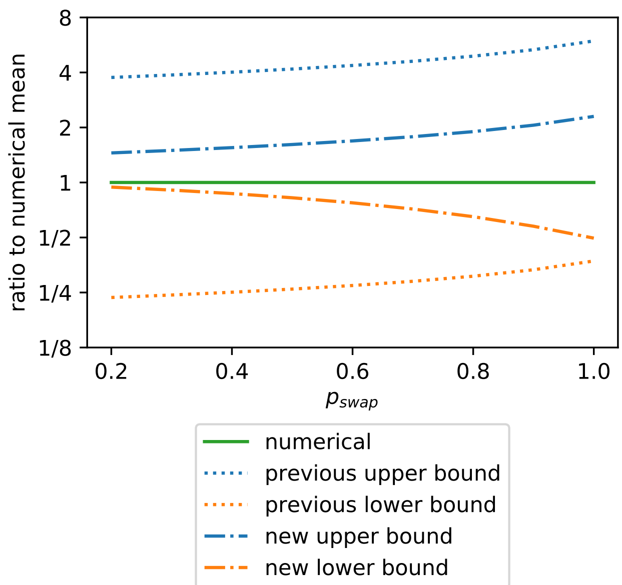


FIG. 4: The ratio of different upper and lower bounds on the mean completion time of a nested repeater protocol, as compared to the numerically calculated mean with the deterministic algorithm from [20], for a repeater chain with 17 nodes ($p_{\rm gen} = 0.5$, entanglement generation is performed in discrete attempts). The figure shows bounds known before this work (eq. (7)) and the tighter bounds from this work in Prop. 3(a) and (b).

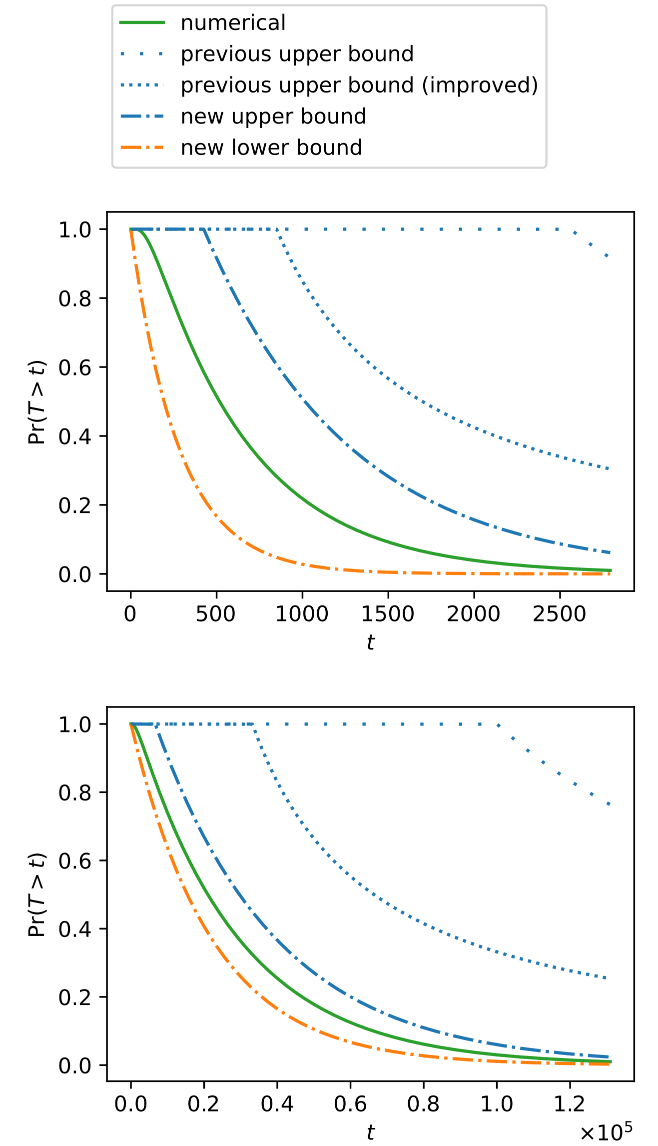


FIG. 5: Probability distribution of the completion time T of a nested repeater protocol. The figure shows the numerically computed distributions using the deterministic algorithm from [20], a polynomially-decaying bound known before this work which is derived from Markov's inequality and a bound on the mean completion time (eq. (9)), and two improvements on eq. (9) we achieve in this work: first, a simple improvement by using Markov's inequality and the improved bound on the mean completion time (Prop. 4(a)), followed by the exponentially-decaying two-sided tail bounds from Prop. 3(c). The plot shows results for a repeater chain with 17 nodes ($p_{\text{gen}} = 0.1$) where entanglement generation is performed in discrete attempts. The swap success probability is $p_{\text{swap}} = 0.5$ (top), and $p_{\text{swap}} = 0.2$ (bottom).

We finish this section by noting a stronger two-sided bound on the completion time T of an equally-spaced repeater chain than Prop. 3(a-b) in the case of deterministic swapping ($p_{\text{swap}} = 1$). The number of segments can be any integer $N \geq 2$. Since we assume that the entanglement swaps take no time (Sec. II A), the mean completion time for this scenario is

$$E[T] = E[\max(T_{\text{gen}}^{(1)}, T_{\text{gen}}^{(2)}, \ldots, T_{\text{gen}}^{(N)})]$$

where $T_{\text{gen}}^{(k)}$ is an independent and identically distributed copy of T_{gen} and describes the completion time of entanglement generation over the k^{th} segment. By replacing $T_{\text{gen}} \to T_{\text{gen}}^{\text{approx}}$, i.e. assuming that the completion time of entanglement generation follows the exponential distribution with mean $1/p_{\text{gen}}$, the following approximation to $E[T]$ has been derived [19, 27]:

$$E[T] \approx \frac{1}{p_{\text{gen}}} \cdot H_N \tag{10}$$

where

$$H_N := \sum_{k=1}^{N} \frac{1}{k} = \gamma + \log(N) + O\left(\frac{1}{N}\right) \tag{11}$$

is the N-th harmonic number and $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. An alternative to eq. (10) is to replace $T_{\text{gen}} \to T_{\text{exp}}$, where T_{exp} is the exponentially-distributed random variable from Lemma 2, which results into

$$E[T] \approx \frac{-1}{\log(1 - p_{\text{gen}})} \cdot H_N = \left(\frac{1}{p_{\text{gen}}} - \frac{1}{2} + O(p_{\text{gen}})\right) \cdot H_N. \tag{12}$$

We remark that eq. (10) and eq. (12) only differ slightly and that their ratio goes to 1 in the limit of $p_{\text{gen}} \to 0$. The quality of the second approximation, eq. (12), has been bounded in work by Eisenberg [54] and to our knowledge no-one has so far noted it in the context of completion times of quantum network protocols. We state it below.

Proposition 4. *[54] Suppose that entanglement swapping is deterministic ($p_{\text{swap}} = 1$). Let $E[T]$ denote the mean completion time of a repeater chain over N segments. Then $E[T]$ is bounded as*

$$a \cdot H_N \leq E[T] \leq 1 + a \cdot H_N$$

where H_N is the N-th harmonic number given in eq. (11) *and*

$$a = \mu_{\text{gen}}^{\text{upper}} - 1 = \frac{-1}{\log(1 - p_{\text{gen}})} = \frac{1}{p_{\text{gen}}} - \frac{1}{2} + O(p_{\text{gen}}).$$

V. SECOND APPLICATION: A QUANTUM SWITCH

Here, we apply our results to a quantum switch. A quantum switch serves k user nodes. Each user is connected to the switch by an arm, which produces bipartite entanglement (a link) between switch and user. As soon as each user has produced a link with the switch, the switch performs a k-fuse operation, i.e. a probabilistic operation converting k bipartite links into a single k-partite entangled state on the user nodes.

Vardoyan et al., considered the scenario in which each user produces entanglement continuously with the switch and the switch fuses whenever it can [10]. They obtained

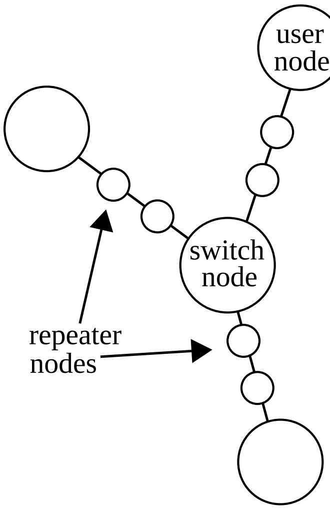


FIG. 6: A quantum switch with 3 users, each connected to the switch by an identical repeater chain which produces links between user and switch. The switch produces 3-partite entangled states, shared between the users, by performing a probabilistic operation on 3 links, one with each user node, as soon as these 3 links are available.

analytical expressions for the rate at which the switch produces multipartite entanglement in the steady-state regime. Here, we consider the alternative protocol where the goal is to produce only a single k-partite state. We go beyond the model of Vardoyan et al., by replacing the arms, which connect the switch to the user, by an arbitrary entanglement-distribution network whose completion time is NBU. An example choice for such a network is the symmetric repeater chain from Sec.IV, yielding the network topology as depicted in Fig. 6, Our tools allow us to achieve bounds on the completion time of the switch, as described in the following proposition.

Proposition 5. *Consider a k-armed quantum switch with fusion success probability p_{fuse}. Suppose that the completion times of the different arms are independent and identically distributed according to an NBU random variable S. Denote by T the time until the switch performs the first successful k-fuse attempt. Then:*

(a) T is NBU;

(b) The mean of T is bounded as

$$E[T] \leq \left(k - 1 + \frac{1}{k}\right) \cdot \frac{E[S]}{p_{\text{fuse}}}.$$

(c) T's tail decays exponentially fast:

$$\Pr(T > t) \leq \exp\left(p_{\text{fuse}} - \frac{p_{\text{fuse}} \cdot t}{(k-1+1/k) \cdot E[S]}\right).$$

Prop. 5(a) follows directly from Prop. 2(a) (Sec. III). Prop. 5(b) is a consequence of the expression for the mean completion time in Prop. 2(b) and the upper bound in Prop. 2(f), while Prop. 5(c) is an instantiation of the tail bound of Prop. 2(c) combined with the mean upper bound of Prop. 5(b).

VI. PROOFS OF MAIN RESULTS

In this section, we prove our main results from Sec. III. We provide proofs in the following order. First, a proof of Lemma 2. Then, we will prove Prop. 2. Since Prop. 1 is a special case of Prop. 2, we do not prove it separately.

A. Proof of Lemma 2

Here, we prove the four parts of Lemma 2: (i) that T_{gen}, the completion time of heralded entanglement generation with probability p_{gen}, is stochastically dominated by $T_{\text{gen}}^{\text{upper}} = 1 + T_{\text{exp}}$, where T_{exp} is exponentially distributed with parameter $-1/\log(1-p_{\text{gen}})$. Next, (ii) that the mean of $T_{\text{gen}}^{\text{upper}}$ equals

$$1 - \frac{1}{\log(1-p_{\text{gen}})} = \frac{1}{p_{\text{gen}}} + \frac{1}{2} + O(p_{\text{gen}}).$$

Then, (iii) that $0 \leq E[T_{\text{gen}}^{\text{upper}}] - E[T_{\text{gen}}] \leq \frac{1}{2}$ and (iv) that $0 \leq E[T_{\text{gen}}^{\text{upper}}]/E[T_{\text{gen}}] \leq 1 + p_{\text{gen}}/2$. Fifth, (v) that $T_{\text{gen}}^{\text{upper}}$ is NBU.

Regarding (i), we use the definition of the geometric distribution in eq. (1), from which it follows that the survival function of T_{gen} is given by

$$\Pr(T_{\text{gen}} > t) = (1-p_{\text{gen}})^{\lfloor t \rfloor}$$

for all $t \geq 1$, where $\lfloor t \rfloor$ denotes the floor of t: $\lfloor t \rfloor = t$ if t is an integer and it equals the largest integer strictly smaller than t otherwise. For $0 \leq t < 1$, we have $\Pr(T_{\text{gen}} > t) = 1 = \Pr\big(T_{\text{gen}}^{\text{upper}} > t\big)$, so the definition of stochastic dominance (Def. 2) is trivially satisfied on the interval $t \in [0, 1)$. We therefore only need to consider $t \geq 1$. Using the notation from Lemma 2, we now bound

$$\begin{aligned}
\Pr(T_{\text{gen}} > t) &= (1-p_{\text{gen}})^{\lfloor t \rfloor} \\
&\leq (1-p_{\text{gen}})^{t-1} \\
&= \exp\left[(t-1) \cdot \log(1-p_{\text{gen}})\right] \\
&\stackrel{*}{=} \Pr(T_{\text{exp}} > t-1) \\
&= \Pr(1 + T_{\text{exp}} > t),
\end{aligned}$$

where in *, we have used the definition of the exponential distribution from eq. (2). For proving (ii), we recall that the mean of an exponential distribution with co-CDF $e^{-\lambda t}$ with parameter $\lambda > 0$ is $1/\lambda$, hence the mean of $T_{\text{gen}}^{\text{upper}}$ is

$$\begin{aligned}
E[T_{\text{gen}}^{\text{upper}}] &= E[1 + T_{\text{exp}}] \\
&= 1 + E[T_{\text{exp}}] \\
&= 1 - \frac{1}{\log(1-p_{\text{gen}})} \\
&= \frac{1}{p_{\text{gen}}} + \frac{1}{2} + O(p_{\text{gen}})
\end{aligned}$$

where in the last equation, we used the expansion of $1/\log(1+x)$ for $|x| < 1$ by Kowalenko [55]. We show (iii) by computing the derivative of $E[T_{\text{gen}}^{\text{upper}}] - E[T_{\text{gen}}]$ as function of p_{gen}, which equals

$$\frac{-1}{(1-p_{\text{gen}})\log^2(1-p_{\text{gen}})} + \frac{1}{p_{\text{gen}}^2}. \tag{13}$$

It is not hard to see that eq. (13) is upper bounded by 0 for all $p_{\text{gen}} \in (0,1)$: we start with the well-established inequality[56]

$$\log(x) \geq \frac{x-1}{\sqrt{x}}$$

for $0 < x \leq 1$, which after the substitution $x \to 1 - p_{\text{gen}}$ becomes

$$\log(1-p_{\text{gen}}) \geq \frac{-p_{\text{gen}}}{\sqrt{1-p_{\text{gen}}}}. \tag{14}$$

Since both sides of eq. (14) are negative and the squaring function $x \mapsto x^2$ is monotonically decreasing for $x \leq 0$, squaring both sides requires the inequality sign to flip,

$$\log^2(1-p_{\text{gen}}) \leq \frac{p_{\text{gen}}^2}{1-p_{\text{gen}}}$$

and hence $(1-p_{\text{gen}})\log^2(1-p_{\text{gen}}) \leq p_{\text{gen}}^2$, implying that the derivative in eq. (13) is upper bounded by 0 for all $p_{\text{gen}} \in (0,1)$. Therefore, $E[T_{\text{gen}}^{\text{upper}}] - E[T_{\text{gen}}]$ is monotonically decreasing in that regime and achieves its optima at $p_{\text{gen}} \downarrow 0$ and $p_{\text{gen}} \uparrow 1$, which are $\frac{1}{2}$ and 0, respectively, yielding precisely the bound in (iii). For showing (iv), divide each side of $0 \leq E[T_{\text{gen}}^{\text{upper}}] - E[T_{\text{gen}}] \leq \frac{1}{2}$ by $E[T_{\text{gen}}]$ to obtain

$$0 \leq \frac{E[T_{\text{gen}}^{\text{upper}}]}{E[T_{\text{gen}}]} - 1 \leq \frac{1}{2E[T_{\text{gen}}]} = \frac{p_{\text{gen}}}{2}$$

from which (iv) directly follows. For proving (v), that $T_{\text{gen}}^{\text{upper}} = 1 + T_{\text{exp}}$ is an NBU random variable, we consider two cases with respect to the definition of NBU (Def. 1):

- both $x < 1$ and $y < 1$. Then

  $$\Pr(1 + T_{\text{exp}} > x) = \Pr(1 + T_{\text{exp}} > y) = 1$$

  so the definition of NBU trivially holds by the fact that $\Pr(1 + T_{\text{exp}} > x + y)$ cannot exceed 1;

- at least one of x or y is 1 or larger. Assume without loss of generality that $y \geq 1$. Then note that $\Pr(1 + T_{\text{exp}} > x + y)$ equals

  $$\begin{aligned} &\Pr(T_{\text{exp}} > x + (y-1)) \\ &\leq \Pr(T_{\text{exp}} > x)\Pr(T_{\text{exp}} > y-1) \\ &= \Pr(T_{\text{exp}} > x)\Pr(1 + T_{\text{exp}} > y) \end{aligned}$$

  where the inequality holds by the fact that T_{exp} is itself NBU (see Example 2). The proof finishes by noting that $1 + T_{\text{exp}}$ stochastically dominates T_{exp}, i.e. $\Pr(1 + T_{\text{exp}} > y) \geq \Pr(T_{\text{exp}} > y)$.

B. Proof of Proposition 2

Now, we prove Prop. 2, which automatically proves its special case Prop. 1. For our proof, we first give a formal definition of T_{output}, following Brand et al. [20]. The RESTART-UNTIL-SUCCESS acts on n quantum states, which first need to have been delivered. Thus, we define a fresh random variable to refer to the time until the last of n quantum states has been delivered:

$$M := \max(T_1, \ldots, T_n).$$

The restarts of the RESTART-UNTIL-SUCCESS protocol, according to a constant success probability p, result in the fact that T_{output} can be written as a *geometric sum* of copies of M:

$$T_{\text{output}} = \sum_{k=1}^{K} M^{(k)} \tag{15}$$

where $M^{(k)}$ is an i.i.d. copy of M and K is a geometrically distributed random variable with parameter p:

$$\Pr(K = k) = p(1-p)^{k-1}. \tag{16}$$

Eq. (15) reflects the fact that the RESTART-UNTIL-SUCCESS protocol needs to perform K attempts at success, each of which takes time given by a fresh instance of M (for a more thorough explanation, see [20]).

Now we will prove each of the statements (a-f) from Prop. 2. For statement (a), we need to show that T_{output} is NBU. This follows directly from the following two facts:

(i) NBU-ness is preserved under the maximum: if $T_1, \ldots, T_n$ are NBU random variables, then so is M;

(ii) NBU-ness is preserved under the geometric sum: if M is an NBU random variables, then so is $T_{\text{output}} = \sum_{k=1}^{K} M^{(k)}$.

We prove item (i) in Appendix A, while item (ii) was proven by Brown, see Sec. 3.2 in [57] [1].

For proving statement (b), $E[T_{\text{output}}] = m/p$ with $m = E[M]$, we apply a well-known fact of randomized sums called Wald's Lemma [58] to eq. (15), which results in

$$E[T_{\text{output}}] = E[M] \cdot E[K]$$

and hence $E[T_{\text{output}}] = m \cdot \frac{1}{p}$.

Statement (c) describes a two-sided bound on the co-CDF of T_{output}:

$$\exp\left(\frac{-p \cdot t}{m} \cdot \frac{1}{1-p}\right) \leq \Pr(T_{\text{output}} > t) \leq \exp\left(p - \frac{p \cdot t}{m}\right).$$

These bounds follow from the following lemma from Brown, see eq.3.2.4 in [57]:

Lemma 3. *[57] Let X be a real-valued random variable with* $\Pr(X < 0) = 0$. *Define the geometric compound sum*

[1] Let us clarify here that the work by Brown proves that the NBU property is preserved under the geometric sum if K is distributed according to eq. (16). However, the same paper also proves that if K is shifted by 1, i.e. $\Pr(K = k) = p(1-p)^k$, then the geometric sum is *always* NWU, irrespective of the summand random variable. However, we will not use the latter case here.

of i.i.d. copies of X as $Y := \sum_{k=1}^{K} X^{(k)}$, where K follows the geometric distribution with success probability p (eq. (16)). Moreover, define $Y_0 := \sum_{k=1}^{K_0} X^{(k)}$, where $K_0 = K - 1$. Then

$$\Pr(Y > t) \leq \exp(p) \exp\left(-t/E[Y]\right)$$

while

$$\Pr(Y > t) \geq \exp\left(-t/E[Y_0]\right).$$

Now interpret $Y \to T_{\text{output}}$ and $X \to M$ in Lemma 3. The upper bound in statement (c) follows directly from Lemma 3 by the use of statement (b), which says that $E[T_{\text{output}}] = m/p$, while for the lower bound in statement (c) we use

$$\begin{aligned} E[Y_0] &= E[K_0] \cdot E[X] \\ &= E[K_0] \cdot E[M] \\ &= \left(\frac{1}{p} - 1\right) \cdot m \\ &= (1-p) \cdot \frac{m}{p}. \end{aligned}$$

Next, (d) states that $T_{\text{output}}/E[T_{\text{output}}]$ approaches the exponential distribution with mean 1. For proving this statement, we substitute $t \to t \cdot E[T_{\text{output}}] = tm/p$ in statement (c). The result is a bound on

$$\Pr(T_{\text{output}} > t \cdot E[T_{\text{output}}]) = \Pr(T_{\text{output}}/E[T_{\text{output}}] > t)$$

given by

$$\exp\left(-t \cdot \frac{1}{1-p}\right) \leq \Pr(T_{\text{output}}/E[T_{\text{output}}] > t) \leq \exp\left(p - t\right).$$

Letting $p \to 0$, the bounds on both sides coincide, and thus

$$\lim_{p \to 0} \Pr(T_{\text{output}}/E[T_{\text{output}}] > t) = \exp\left(-t\right)$$

which is precisely the co-CDF of the exponential distribution with parameter 1.

For showing the upper bound in statement (e),

$$m \leq \sum_{j=1}^{n} E[T_j]$$

we use the fact that for all $j = 1, \ldots, n$, it holds that $T_j \geq 0$. The maximum of of nonnegative numbers is upper bounded by its sum, and thus

$$\begin{aligned} m &= E[\max(T_1, \ldots, T_n)] \\ &= \sum_{t_1,\ldots,t_n} \Pr(T_1 = t_1, \ldots, T_n = t_n) \max(t_1, \ldots, t_n) \\ &\leq \sum_{t_1,\ldots,t_n} \Pr(T_1 = t_1, \ldots, T_n = t_n)\,(t_1 + \cdots + t_n) \\ &\stackrel{*}{=} \sum_{j=1}^{n} \sum_{t_j} \Pr(T_j = t_j) t_j \\ &= E\left[\sum_{j=1}^{n} T_j\right] \end{aligned}$$

where for $*$ we made use of the fact that all T_j are independent. The proof for the lower bound in statement (e), $\max_{1\leq j\leq n} E[T_j] \leq m$, is similar and relies on the fact that $\max(t_1, \ldots, t_n) \geq t_j$ for all $1 \leq j \leq n$, where $t_1, \ldots, t_n$ are nonnegative numbers. Last, (f) states that if all T_j are identically distributed with mean $E[T]$, then

$$1 \leq \frac{m}{E[T]} \leq n - 1 + \frac{1}{n}$$

where we recall that $m = E[\max(T_1, \ldots, T_n)]$. For proving this statement, we need the following lemma from Hu and Lin [59, Lemma 2.2.].

Lemma 4. *[59] If $X_1, \ldots, X_n$ are independent and identically distributed copies of an NBU random variable X on the domain $[0, \infty)$, then $E[\min(X_1, \ldots, X_n)] \geq E[X]/n$.*

Proof. The proof is based on two facts. First, note that

$$\Pr(\min(X_1, \ldots, X_n) > x) = \prod_{j=1}^{n} \Pr(X_j > x) = \Pr(X > x)^n.$$

Second, note that if X is NBU, then by repeated application of the definition of NBU (Def. 1), we find that

$$\Pr\left(X > \sum_{j=1}^{n} x_j\right) \leq \prod_{j=1}^{n} \Pr(X > x_j)$$

for any nonnegative numbers $x_j, 1 \leq j \leq n$. When choosing all x_j identical, say, to some constant nonnegative number x, this reduces to

$$\Pr(X > nx) \leq \Pr(X > x)^n.$$

Using these two facts, we can now prove the lemma:

$$\begin{aligned} E[\min(X_1, \ldots, X_n)] &= \int_0^\infty \Pr(X > x)^n dx \\ &\geq \int_0^\infty \Pr(X > nx) dx \\ &= \int_0^\infty \Pr(X/n > x) dx \\ &= E[X/n] \\ &= E[X]/n \end{aligned}$$

where we have used the fact that for any real-valued random variable X with $\Pr(X < 0) = 0$, the mean can be computed as $E[X] = \int_0^\infty \Pr(X > x)dx$. □

Statement (f) is proven by noting that for nonnegative numbers $t_1, \ldots, t_n$, it holds that $t_j \geq \min(t_1, \ldots, t_n)$ for all $j = 1, \ldots, n$, and therefore

$$t_1 + \ldots t_n \geq \max(t_1, \ldots, t_n) + (n-1) \cdot \min(t_1, \ldots, t_n).$$

Translating this to the T_j yields

$$\begin{aligned} E\left[\sum_{j=1}^{n} T_j\right] \geq (n-1) \cdot E[\min(T_1, \ldots, T_n] \\ + E[\max(T_1, \ldots, T_n)]. \qquad (17) \end{aligned}$$

The left hand side of eq. (17) equals $n \cdot E[T]$ by the fact that the T_j are i.i.d., while the right hand side is lower bounded by $(n-1)/n \cdot E[T] + E[\max(T_1, \ldots, T_n)]$ by Lemma 4. Reshuffling yields

$$\begin{aligned} E[\max(T_1, \ldots, T_n) &\leq n \cdot E[T] - \frac{n-1}{n} E[T] \\ &= \left(n - 1 + \frac{1}{n}\right) E[T]. \end{aligned}$$

which is what we set out to prove.

VII. DISCUSSION

The distribution of remote entanglement is a key element of many quantum network applications. In this work, we provided analytical bounds on both the mean and quantiles of entanglement delivery times for a large class of protocols. We applied these results to a nested quantum repeater chain scheme and to a quantum switch, and obtained bounds which are tighter than present in the literature.

In particular, we considered a frequently-used approximation to the mean entanglement-delivery time in the nested repeater chain scheme, known as the 3-over-2 formula. This approximation is derived by assuming that the delivery time follows an exponential distribution at each nesting level. It was not known in general how good this approximation is. Moreover, finding the exact mean delivery time has been an open problem for more than ten years [25]. We made a large step towards solving this question by showing that the co-CDF of the delivery time, i.e. the probability that entanglement is delivered after time t, is lower bounded by the co-CDF of an exponential distribution, and upper bounded by the co-CDF of an exponential distribution multiplied by a factor which is independent of t. In the limit of small success probabilities of the repeater's components, the bounds coincide. Second, we show that the 3-over-2 formula is, in essence, an upper bound to the mean delivery time, rendering old analyses building upon this approximation pessimistic.

Regarding future work, note that in many quantum internet scenarios, already-produced entanglement waits for the generation of other entanglement and in the meantime suffers from memory noise. We leave for future work converting our bounds on the delivery time to bounds on the amount of memory noise, and thus on the quality of the produced state.

In this work we only focused on the first remote entanglement that is delivered. Some protocols, however, might deliver entanglement while still holding residual entanglement, for example at lower levels in case of the nested repeater chain. In such a case, it is not optimal to restart the protocol for producing a second entangled pair of qubits, since that would require discarding the residual entanglement. Hence, another possibility for future work would be to extend our results to protocols which produce multiple entangled pairs without discarding existing entanglement in between.

Our bounds are partially based on a novel connection with reliability theory. We expect that reliability-theoretic tools will be useful in solving other open problems in quantum networks too.

ACKNOWLEDGEMENTS

The authors would like to thank Kenneth Goodenough, Boxi Li and Filip Rozpędek for helpful discussions, and would like to thank Kenneth Goodenough, Boxi Li and Gayane Vardoyan for critical reading of the manuscript. This work was supported by the QIA project (funded by European Union's Horizon 2020, Grant Agreement No. 820445), the NEASQC project (Horizon 2020, Grant Agreement No. 951821) and by the Netherlands Organization for Scientific Research (NWO/OCW), as part of the Quantum Software Consortium program (project number 024.003.037 / 3368).

[1] H. J. Kimble, "The quantum internet," *Nature*, vol. 453, no. 7198, pp. 1023–1030, Jun 2008. [Online]. Available: https://doi.org/10.1038/nature07127

[2] S. Wehner, D. Elkouss, and R. Hanson, "Quantum internet: A vision for the road ahead," *Science*, vol. 362, no. 6412, 2018. [Online]. Available: https://science.sciencemag.org/content/362/6412/eaam9288

[3] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Proceedings of IEEE International Conference on Computers, Systems and Signal Processing*, vol. 175, 1984.

[4] A. K. Ekert, "Quantum cryptography based on Bell's theorem," *Phys. Rev. Lett.*, vol. 67, pp. 661–663, Aug 1991. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.67.661

[5] A. M. Childs, "Secure assisted quantum computation," *Quantum Info. Comput.*, vol. 5, no. 6, pp. 456–466, Sep. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=2011670.2011674

[6] A. Kellerer, "Quantum telescopes," *Astronomy & Geophysics*, vol. 55, no. 3, pp. 3.28–3.32, 06 2014. [Online]. Available: https://doi.org/10.1093/astrogeo/atu126

[7] H.-J. Briegel, W. Dür, J. I. Cirac, and P. Zoller, "Quantum repeaters: The role of imperfect local operations in quantum communication," *Phys. Rev. Lett.*, vol. 81, pp. 5932–5935, Dec 1998. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.81.5932

[8] W. J. Munro, K. Azuma, K. Tamaki, and K. Nemoto, "Inside quantum repeaters," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 21, no. 3, pp. 78–90, may 2015. [Online]. Available: https://doi.org/10.1109%2Fjstqe.2015.2392076

[9] S. Muralidharan, L. Li, J. Kim, N. Lütkenhaus, M. D. Lukin, and L. Jiang, "Optimal architectures for long distance quantum communication," *Scientific Reports*, vol. 6, pp. 20 463 EP –, Feb 2016, article. [Online]. Available: http://dx.doi.org/10.1038/srep20463

[10] G. Vardoyan, S. Guha, P. Nain, and D. Towsley, "On the stochastic analysis of a quantum entanglement switch," *SIGMETRICS Perform. Eval. Rev.*, vol. 47, no. 2, pp. 27–29, Dec. 2019. [Online]. Available: https://doi.org/10.1145/3374888.3374899

[11] A. Pirker, J. Wallnöfer, and W. Dür, "Modular architectures for quantum networks," *New Journal of Physics*, vol. 20, no. 5, p. 053054, may 2018. [Online]. Available: https://doi.org/10.1088%2F1367-2630%2Faac2aa

[12] J. Wallnöfer, A. Pirker, M. Zwerger, and W. Dür, "Multipartite state generation in quantum networks with optimal scaling," *Scientific Reports*, vol. 9, no. 1, p. 314, 2019. [Online]. Available: https://doi.org/10.1038/s41598-018-36543-5

[13] 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, p. 25, Mar 2019. [Online]. Available: https://doi.org/10.1038/s41534-019-0139-x

[14] V. V. Kuzmin, D. V. Vasilyev, N. Sangouard, W. Dür, and C. A. Muschik, "Scalable repeater architectures for multi-party states," *npj Quantum Information*, vol. 5, no. 1, p. 115, Dec 2019. [Online]. Available: https://doi.org/10.1038/s41534-019-0230-3

[15] M. Schmitt, J. Redi, P. Cesar, and D. Bulterman, "1Mbps is enough: Video quality and individual idiosyncrasies in multiparty HD video-conferencing," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[16] N. Kalb, P. C. Humphreys, J. J. Slim, and R. Hanson, "Dephasing mechanisms of diamond-based nuclear-spin memories for quantum networks," *Phys. Rev. A*, vol. 97, p. 062330, Jun 2018. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.97.062330

[17] M. F. Askarani, T. Lutz, M. G. Puigibert, N. Sinclair, D. Oblak, and W. Tittel, "Persistent atomic frequency comb based on Zeeman sub-levels of an erbium-doped crystal waveguide," *J. Opt. Soc. Am. B*, vol. 37, no. 2, pp. 352–358, Feb 2020. [Online]. Available: http://josab.osa.org/abstract.cfm?URI=josab-37-2-352

[18] R. V. Meter, T. D. Ladd, W. J. Munro, and K. Nemoto, "System design for a long-line quantum repeater," *IEEE/ACM Transactions on Networking*, vol. 17, no. 3, pp. 1002–1013, June 2009.

[19] E. Shchukin, F. Schmidt, and P. van Loock, "Waiting time in quantum repeaters with probabilistic entanglement swapping," *Phys. Rev. A*, vol. 100, p. 032322, Sep 2019. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.100.032322

[20] S. Brand, T. Coopmans, and D. Elkouss, "Efficient computation of the waiting time and fidelity in quantum repeater chains," *IEEE Journal on Selected Areas in Communications*, pp. 619 – 639, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8972391

[21] B. Li, T. Coopmans, and D. Elkouss, "Efficient optimization of cut-offs in quantum repeater chains," *arXiv:2005.04946*, 2020.

[22] M. Caleffi, "Optimal routing for quantum networks," *IEEE Access*, vol. 5, pp. 22 299–22 312, 2017.

[23] S. E. Vinay and P. Kok, "Statistical analysis of quantum-entangled-network generation," *Phys. Rev. A*, vol. 99, p. 042313, Apr 2019. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.99.042313

[24] S. Khatri, C. T. Matyas, A. U. Siddiqui, and J. P. Dowling, "Practical figures of merit and thresholds for entanglement distribution in quantum networks," *Phys. Rev. Research*, vol. 1, p. 023032, Sep 2019. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevResearch.1.023032

[25] N. Sangouard, C. Simon, H. de Riedmatten, and N. Gisin, "Quantum repeaters based on atomic ensembles and linear optics," *Rev. Mod. Phys.*, vol. 83, pp. 33–80, Mar 2011. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.83.33

[26] V. V. Kuzmin and D. V. Vasilyev, "Diagrammatic technique for simulation of large-scale quantum repeater networks with dissipating quantum memories," *arXiv:2009.10415*, 2020.

[27] F. Schmidt and P. van Loock, "Memory-assisted long-distance phase-matching quantum key distribution," *Phys. Rev. A*, vol. 102, p. 042614, Oct 2020. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.102.042614

[28] O. A. Collins, S. D. Jenkins, A. Kuzmich, and T. A. B. Kennedy, "Multiplexed memory-insensitive quantum repeaters," *Phys. Rev. Lett.*, vol. 98, p. 060502, Feb 2007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.98.060502

[29] N. K. Bernardes, L. Praxmeyer, and P. van Loock, "Rate analysis for a hybrid quantum repeater," *Phys. Rev. A*, vol. 83, p. 012323, Jan 2011. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.83.012323

[30] L. Praxmeyer, "Reposition time in probabilistic imperfect memories," *arXiv:1309.3407*, 2013. [Online]. Available: https://arxiv.org/abs/1309.3407

[31] J. Calsamiglia and N. Lütkenhaus, "Maximum efficiency of a linear-optical bell-state analyzer," *Applied Physics B*, vol. 72, no. 1, pp. 67–71, Jan 2001. [Online]. Available: https://doi.org/10.1007/s003400000484

[32] S. Santra, L. Jiang, and V. S. Malinovsky, "Quantum repeater architecture with hierarchically optimized memory buffer times," *Quantum Science and Technology*, vol. 4, no. 2, p. 025010, mar 2019. [Online]. Available: https://doi.org/10.1088%2F2058-9565%2Fab0bc2

[33] K. Chakraborty, F. Rozpedek, A. Dahlberg, and S. Wehner, "Distributed routing in a quantum internet," *arXiv:1907.11630*, 2019.

[34] K. Azuma, S. Bäuml, T. Coopmans, D. Elkouss, and B. Li, "Tools for quantum network design," *AVS Quantum Science*, vol. 3, no. 1, p. 014101, 2021. [Online]. Available: https://doi.org/10.1116/5.0024062

[35] L.-M. Duan, M. D. Lukin, J. I. Cirac, and P. Zoller, "Long-distance quantum communication with atomic ensembles and linear optics," *Nature*, vol. 414, pp. 413 EP –, Nov 2001, article. [Online]. Available: https://doi.org/10.1038/35106500

[36] L. Jiang, J. M. Taylor, and M. D. Lukin, "Fast and robust approach to long-distance quantum communication with atomic ensembles," *Phys. Rev. A*, vol. 76, p. 012301, Jul 2007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.76.012301

[37] C. Simon, H. de Riedmatten, M. Afzelius, N. Sangouard, H. Zbinden, and N. Gisin, "Quantum repeaters with photon pair sources and multimode memories," *Phys. Rev. Lett.*, vol. 98, p. 190503, May 2007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.98.190503

[38] J. B. Brask and A. S. Sørensen, "Memory imperfections in atomic-ensemble-based quantum repeaters," *Phys. Rev. A*, vol. 78, p. 012350, Jul 2008. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.78.012350

[39] N. Sangouard, C. Simon, J. c. v. Minář, H. Zbinden, H. de Riedmatten, and N. Gisin, "Long-distance entanglement distribution with single-photon sources," *Phys. Rev. A*, vol. 76, p. 050301, Nov 2007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.76.050301

[40] N. Sangouard, C. Simon, B. Zhao, Y.-A. Chen, H. de Riedmatten, J.-W. Pan, and N. Gisin, "Robust and efficient quantum repeaters with atomic ensembles and linear optics," *Phys. Rev. A*, vol. 77, p. 062301, Jun 2008. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.77.062301

[41] N. Sangouard, R. Dubessy, and C. Simon, “Quantum repeaters based on single trapped ions,” *Phys. Rev. A*, vol. 79, p. 042340, Apr 2009. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.79.042340

[42] S. Abruzzo, S. Bratzik, N. K. Bernardes, H. Kampermann, P. van Loock, and D. Bruß, “Quantum repeaters and quantum key distribution: Analysis of secret-key rates,” *Phys. Rev. A*, vol. 87, p. 052315, May 2013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.87.052315

[43] K. Boone, J.-P. Bourgoin, E. Meyer-Scott, K. Heshami, T. Jennewein, and C. Simon, “Entanglement over global distances via quantum repeaters with satellite links,” *Phys. Rev. A*, vol. 91, p. 052325, May 2015. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.91.052325

[44] F. Kimiaee Asadi, N. Lauk, S. Wein, N. Sinclair, C. O’Brien, and C. Simon, “Quantum repeaters with individual rare-earth ions at telecommunication wavelengths,” *Quantum*, vol. 2, p. 93, Sep. 2018. [Online]. Available: https://doi.org/10.22331/q-2018-09-13-93

[45] N. Lo Piparo, W. J. Munro, and K. Nemoto, “Quantum multiplexing,” *Phys. Rev. A*, vol. 99, p. 022337, Feb 2019. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.99.022337

[46] F. K. Asadi, S. C. Wein, and C. Simon, “Long-distance quantum communication with single ^{167}er ions,” 2020.

[47] K. Sharman, F. K. Asadi, S. C. Wein, and C. Simon, “Quantum repeaters based on individual electron spins and nuclear-spin-ensemble memories in quantum dots,” *arXiv:2010.13863*, 2020.

[48] Y. Wu, J. Liu, and C. Simon, “Near-term performance of quantum repeaters with imperfect ensemble-based quantum memories,” *Phys. Rev. A*, vol. 101, p. 042301, Apr 2020. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.101.042301

[49] C. Liorni, H. Kampermann, and D. Bruss, “Quantum repeaters in space,” *arXiv:2005.10146*, 2020.

[50] N. H. Nickerson, Y. Li, and S. C. Benjamin, “Topological quantum computing with a very noisy network and local error rates approaching one percent,” *Nature Communications*, vol. 4, no. 1, apr 2013. [Online]. Available: https://doi.org/10.1038%2Fncomms2773

[51] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, “Purification of noisy entanglement and faithful teleportation via noisy channels,” *Phys. Rev. Lett.*, vol. 76, pp. 722–725, Jan 1996. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.76.722

[52] D. Deutsch, A. Ekert, R. Jozsa, C. Macchiavello, S. Popescu, and A. Sanpera, “Quantum privacy amplification and the security of quantum cryptography over noisy channels,” *Phys. Rev. Lett.*, vol. 77, pp. 2818–2821, Sep 1996. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.77.2818

[53] A. W. Marshall and I. Olkin, *Nonparametric Families: Origins in Reliability Theory.* New York, NY: Springer New York, 2007, pp. 137–193. [Online]. Available: https://doi.org/10.1007/978-0-387-68477-2_5

[54] B. Eisenberg, “On the expectation of the maximum of IID geometric random variables,” vol. 78, no. 2, pp. 135–143, 2008, 135.

[55] V. Kowalenko, “Properties and applications of the reciprocal logarithm numbers,” *Acta Applicandae Mathematicae*, vol. 109, no. 2, pp. 413–437, oct 2008. [Online]. Available: https://doi.org/10.1007%2Fs10440-008-9325-0

[56] F. Topsøe, “Some bounds for the logarithmic function,” *Inequality theory and applications*, vol. 4, no. 01, 2007.

[57] M. Brown, “Error bounds for exponential approximations of geometric convolutions,” *Ann. Probab.*, vol. 18, no. 3, pp. 1388–1402, 07 1990. [Online]. Available: https://doi.org/10.1214/aop/1176990750

[58] A. Wald, “Sequential analysis,” 1947.

[59] C.-Y. Hu and G. D. Lin, “Characterizations of the exponential distribution by stochastic ordering properties of the geometric compound,” *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 3, pp. 499–506, sep 2003. [Online]. Available: https://doi.org/10.1007%2Fbf02517803

Appendix A: Proof that the NBU property is preserved under the maximum

Here, we prove that the NBU property is preserved under the maximum of independent random variables.

Lemma 5. *Suppose $X_1, \ldots, X_n$ are independent random variables (not necessarily identically distributed). If all X_j are NBU random variables, then so is* $\max(X_1, \ldots, X_n)$.

We first prove the special case for $n = 2$, from which the statement for general n follows.

Lemma 6. *Let A and B be independent nonnegative real-valued random variables (not necessarily identically distributed). If both are NBU, then so is* $\max(A, B)$.

Proof. Let us denote $a_z := \Pr(A > z)$ and $b_z := \Pr(B > z)$ for $z \geq 0$. Assume that A and B possess the NBU property (Def. 1), so that

$$a_{x+y} \leq a_x a_y \text{ and } b_{x+y} \leq b_x b_y \quad \text{for all } x, y \geq 0. \tag{A1}$$

We also write $m_z := \Pr(\max(A, B) \geq z)$ and compute

$$\begin{aligned} m_z &= \Pr(\max(A, B) > z) \\ &= 1 - \Pr(\max(A, B) \leq z) \\ &= 1 - \Pr(A \leq z) \Pr(B \leq z) \\ &= 1 - (1 - a_z)(1 - b_z) && \text{(A2)} \\ &= a_z + b_z - a_z b_z \\ &= a_z + b_z(1 - a_z). && \text{(A3)} \end{aligned}$$

We will prove that $\max(A,B)$ is NBU, which in our notation becomes $m_{x+y} \leq m_x m_y$ for all $x,y \geq 0$. To begin, we write out the expressions for both sides, i.e. for m_{x+y} and for $m_x m_y$. First, using eq. (A2), we write out

$$m_{x+y} = 1 - (1 - a_{x+y})(1 - b_{x+y}). \tag{A4}$$

Since m_{x+y} from eq. (A4) is monotonically increasing in a_{x+y} and moreover $a_{x+y} \leq a_x a_y$ (eq. (A1)), we obtain

$$m_{x+y} \leq 1 - (1 - a_x a_y)(1 - b_{x+y}). \tag{A5}$$

We use the same insight again, but now for b_{x+y}: the right-hand side of eq. (A5) is monotonically increasing in b_{x+y}, which combined with the fact that $b_{x+y} \leq b_x b_y$ (eq. (A1)) yields

$$m_{x+y} \leq 1 - (1 - a_x a_y)(1 - b_x b_y) = a_x a_y + b_x b_y (1 - a_x a_y). \tag{A6}$$

Next, by eq. (A3) we have

$$\begin{aligned} m_x m_y &= (a_x + b_x(1 - a_x)) \cdot (a_y + b_y(1 - a_y)) \\ &= a_x a_y + a_x b_y (1 - a_y) + a_y b_x (1 - a_x) + b_x b_y (1 - a_x)(1 - a_y). \end{aligned} \tag{A7}$$

In order to prove that $m_{x+y} \leq m_x m_y$ we consider three cases.

- **Case $\boldsymbol{b_x = 0}$.** In this case eq. (A6) reduces to $m_{x+y} \leq a_x a_y$ and eq. (A7) becomes

$$m_x m_y = a_x a_y + a_x b_y (1 - a_y). \tag{A8}$$

  Since a_x, a_y, b_x and b_y are all cumulative probabilities, they take values in the interval $[0,1]$, and therefore the second term of eq. (A8) is nonnegative, which yields $m_x m_y \geq a_x a_y \geq m_{x+y}$.

- **Case $\boldsymbol{b_y = 0}$.** By the fact that both the right hand side of eq. (A6) as well as the expression for $m_x m_y$ (eq. (A7)) are invariant under exchanging b_x and b_y, this case is proven identically to the first case.

- **Case $\boldsymbol{b_x \neq 0}$ and $\boldsymbol{b_y \neq 0}$.** Using eq. (A6) and eq. (A7), we expand

$$\begin{aligned} \frac{m_{x+y} - m_x m_y}{b_x b_y} &= \frac{a_x a_y}{b_x b_y} + \frac{b_x b_y}{b_x b_y}(1 - a_x a_y) - \frac{a_x a_y}{b_x b_y} - \frac{a_x b_y}{b_x b_y}(1 - a_y) - \frac{a_y b_x}{b_x b_y}(1 - a_x) - \frac{b_x b_y}{b_x b_y}(1 - a_x) \cdot (1 - a_y) \\ &= 1 - a_x a_y - \frac{a_x}{b_x}(1 - a_y) - \frac{a_y}{b_y}(1 - a_x) - (1 - a_x) \cdot (1 - a_y) \end{aligned}$$

  Using the fact that $b_x, b_y \leq 1$, we obtain

$$\frac{m_{x+y} - m_x m_y}{b_x b_y} \leq 1 - a_x a_y - a_x (1 - a_y) - a_y (1 - a_x) - (1 - a_x) \cdot (1 - a_y) = 0.$$

  Since b_x and b_y are positive numbers, it follows that $m_{x+y} - m_x m_y \leq 0$. This concludes our proof.

□

Let us now show how Lemma 5 follows from Lemma 6. Let $X_1, \ldots, X_n$ be n NBU independent random variables, for $n \geq 2$. We use induction on n. The case $n = 2$ is proven in Lemma 6. Now suppose Lemma 5 holds for $n = m$ for some $m \geq 2$. We show that Lemma 6 also holds for $n = m + 1$. For this, choose $A = \max(X_1, \ldots, X_m)$ and $B = X_{m+1}$. By assumption, B is NBU, and so is A by the induction hypothesis. Note that

$$\begin{aligned} \max(X_1, \ldots, X_m, X_{m+1}) &= \max(\max(X_1, \ldots, X_m), X_{m+1}) \\ &= \max(A, B), \end{aligned}$$

so it follows from Lemma 6 that $\max(X_1, \ldots, X_{m+1})$ is also NBU, which concludes the proof of Lemma 5.

Appendix B: Proof of the lower bounds in Proposition 3

Here, we prove the two lower bounds in Prop. 3: first, Prop. 3(b), followed by the lower bound on the quantiles from Prop. 3(c).

Throughout the appendix, we will use the notation $X^{(1)}, X^{(2)}, \ldots$ to denote independent and identically distributed copies of a random variable X. Before proving the bounds on the mean and tail of T_n, let us formally define it. Regarding the base case $n = 0$, which describes elementary-link generation between adjacent nodes, we use either of two flavors: we either set $T_0 = T_{\text{gen}}$, i.e. T_0 follows the geometric distribution with parameter p_{gen}, or we set $T_0 = T_{\text{gen}}^{\text{approx}}$, i.e. T_0 follows the exponential distribution with parameter p_{gen}. For each statement about T_n in this appendix, either the statement will hold for both flavors, or it will be clear from the context which of the two flavors is used. Regardless of the choice for $n = 0$, we define T_n for $n > 0$ as

$$T_{n+1} = \sum_{k=1}^{K} M_n^{(k)} \tag{B1}$$

where K is geometrically distributed with parameter p_{swap} and M_n is defined as

$$M_n = \max(T_n^{(1)}, T_n^{(2)}). \tag{B2}$$

Eq. (B1) was given in [20] and can be found by applying eq. (15) to each nesting level of the repeater protocol, where $M = M_n$ in eq. (15) describes the time until the last of two links, each spanning 2^n repeater segments, has been delivered.

1. Proof of Proposition 3(b)

Here, we will prove the lower bound on the mean completion time T_n of the nested repeater protocol on n nesting levels. Informally stated, the insight is that

$$\max\left(\sum_{k=1}^{K^{(1)}} X^{(k)}, \sum_{k=1}^{K^{(2)}} X^{(k)}\right) \geq_{\text{st}} \sum_{k=1}^{\max(K^{(1)},K^{(2)})} X^{(k)} \quad \text{(informal)}$$

i.e. considering sums with independent and identically distributed summands, the maximum of two sums stochastically dominates the "longest" of the two. Since the definition of M_n in eq. (B2) contains the maximum of two such sums, we use this idea to define a new random variable R_n as the "longest" of the two sums; by the insight above, R_n is stochastically dominated by M_n. Using Lemma 1, this stochastic domination can be converted to $E[M_n] \geq E[R_n]$, after which the bound on the mean of T_n as described in Prop. 3(b) follows by noting that $E[T_n] = E[M_n]/p_{\text{swap}}$.

We now give the formal proof, which we divide into three steps. First, we define R_n and compute its mean. Next, we show that $M_n \geq_{\text{st}} R_n$ for all $n > 0$, from which we infer a lower bound on the mean of T_n as third step.

For the first step, we define R_n:

$$R_0 = \max(T_0^{(1)}, T_0^{(2)}),$$
$$R_{n+1} = \sum_{j=1}^{N} R_n^{(j)} \quad \text{for } n \geq 0.$$

Here, $N = \max\left(K^{(1)}, K^{(2)}\right)$ where $K^{(1)}$ and $K^{(2)}$ are both geometrically distributed with parameter p_{swap}. We emphasize that contrary to T_n, the random variable R_n does not correspond to the completion time of a protocol.

The mean of R_n is computed using the following two lemmas.

Lemma 7. *Let $X^{(1)}$ and $X^{(2)}$ be independent and identically distributed random variables with mean $1/p$ for some $0 < p \leq 1$. If both $X^{(1)}$ and $X^{(2)}$ follow a geometric distribution, then*

$$E[\max(X^{(1)}, X^{(2)})] = \frac{3-2p}{p(2-p)}$$

while if they follow an exponential distribution, then

$$E[\max(X^{(1)}, X^{(2)})] = \frac{3}{2p}.$$

Proof. We start with the case that X follows a geometric distribution. Note that $\min(X^{(1)}, X^{(2)})$ is geometrically distributed with parameter $1-(1-p)^2$:

$$\Pr\Big(\min(X^{(1)}, X^{(2)}) > t\Big) = \Pr\Big(X^{(1)} > t\Big)\Pr\Big(X^{(2)} > t\Big) = (1-p)^t \cdot (1-p)^t = (1-p)^{2t} = \left[1-\left(1-(1-p)^2\right)\right]^t$$

for $t = 0, 1, 2, \ldots$. Combined with the fact that $E[\max(X^{(1)}, X^{(2)})] = E[X^{(1)} + X^{(2)} - \min(X^{(1)}, X^{(2)})] = E[X^{(1)}] + E[X^{(2)}] - E[\min(X^{(1)}, X^{(2)})]$, we obtain

$$E[\max(X^{(1)}, X^{(2)})] = \frac{1}{p} + \frac{1}{p} - \frac{1}{1-(1-p)^2} = \frac{3-2p}{p(2-p)}$$

The case of the exponential distribution is analogous, with $\min(X^{(1)}, X^{(2)})$ following the exponential distribution with parameter $2p$. $\square$

Lemma 8. *The mean of R_n is*

$$E[R_n] = \left(\frac{3-2p_{\text{swap}}}{p_{\text{swap}}(2-p_{\text{swap}})}\right)^n \cdot \nu_0 \tag{B3}$$

where ν_0 is defined as follows. If T_0, which describes elementary-link generation between adjacent nodes, follows the geometric distribution with parameter p_{gen}, then

$$\nu_0 = E[R_0] = E[\max(T_0^{(1)}, T_0^{(2)})] = \frac{3-2p_{\text{gen}}}{p_{\text{gen}}(2-p_{\text{gen}})} \tag{B4}$$

while if T_0 follows the exponential distribution with parameter p_{gen}, then

$$\nu_0 = E[R_0] = E[\max(T_0^{(1)}, T_0^{(2)})] = \frac{3}{2p_{\text{gen}}}. \tag{B5}$$

Proof. We use induction on n. The case $n=0$ is treated in Lemma 7 where we set $p = p_{\text{gen}}$. For the induction case, we note that

$$E[R_{n+1}] = E\left[\sum_{j=1}^{N} R_n^{(j)}\right] = E[N]\cdot E[R_n]$$

by Wald's Lemma [58]. Since $N = \max(K^{(1)}, K^{(2)})$ and K is geometrically distributed with parameter p_{swap}, we again invoke Lemma 7 to obtain

$$E[N] = E[\max(K^{(1)}, K^{(2)})] = \frac{3-2p_{\text{swap}}}{p_{\text{swap}}(2-p_{\text{swap}})}.$$

This finishes the proof. □

As second step, we will show that M_n stochastically dominates R_n, for which we need the following two auxiliary lemmas and corollary.

Lemma 9. *Let P and Q be independent real-valued random variables, and P' and Q' i.i.d. copies of P and Q respectively. Then $P \geq_{\text{st}} Q$ implies $\max(P, P') \geq_{\text{st}} \max(Q, Q')$.*

Proof. By definition of $P \geq_{\text{st}} Q$, we have, for all real numbers z, that $\Pr(P > z) \geq \Pr(Q > z)$ and therefore $\Pr(P \leq z) \leq \Pr(Q \leq z)$. Consequently,

$$\Pr(\max(P,P') > z) = 1 - \Pr(\max(P,P') \leq z) = 1 - \Pr(P \leq z)^2 \geq 1 - \Pr(Q \leq z)^2 = \Pr(\max(Q,Q') > z)$$

for all real numbers z, so $\max(P,P') \geq_{\text{st}} \max(Q,Q')$. □

Lemma 10. *Let P and Q be independent, real-valued random variables with identical domain. Then* $\max(P,Q) \geq_{\text{st}} Q$.

Proof. For any real number z, we have

$$\Pr(\max(P,Q) > z) = 1 - \Pr(\max(P,Q) \leq z) = 1 - \Pr(P \leq z)\Pr(Q \leq z) \overset{*}{\geq} 1 - \Pr(Q \leq z) = \Pr(Q > z)$$

where the inequality * holds because $\Pr(P < z) \leq 1$. □

Corollary 1. *Let $A^{(1)}, A^{(2)}, A^{(3)}$ and $A^{(4)}$ be independent and identically distributed random variables with domain $\{1,2,3,\dots\}$. Furthermore, let X, Y and Z be independent and identically distributed random variables with domain $[0,\infty)$. Then*

$$\max\left(\sum_{a=1}^{A^{(1)}} X^{(a)}, \sum_{b=1}^{A^{(2)}} Y^{(b)}\right) \geq_{\text{st}} \sum_{a=1}^{\max\left(A^{(3)},A^{(4)}\right)} Z^{(a)}. \tag{B6}$$

Proof. We note that random sums occur on both sides of eq. (B6), that is, sums whose number of terms is a random variable. We expand both sides of the inequality from the lemma as a weighted sum over instantiations of this random variable. For the left-hand-side, we obtain

$$\Pr\left(\max\left(\sum_{a=1}^{A^{(1)}} X^{(a)}, \sum_{b=1}^{A^{(2)}} Y^{(b)}\right) > y\right) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \Pr\left(A^{(1)} = i\right)\cdot \Pr\left(A^{(2)} = j\right)\cdot C_{ij}^{y}$$

for $y \geq 0$, where we have defined

$$C_{ij}^{y} := \Pr\left(\max\left(\sum_{a=1}^{i} X^{(a)}, \sum_{b=1}^{j} Y^{(b)} \right) > y \right)$$

and for the right-hand-side we get

$$\Pr\left(\sum_{a=1}^{\max\left(A^{(3)}, A^{(4)}\right)} Z^{(a)} > y \right) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \Pr\left(A^{(3)} = i\right) \cdot \Pr\left(A^{(4)} = j\right) \cdot D_{ij}^{y}$$

with

$$D_{ij}^{y} := \Pr\left(\sum_{a=1}^{\max(i,j)} Z^{(a)} > y \right).$$

Given fixed i and j, we define random variables P and Q as follows:

- if $\max(i,j) = i > j$, then define $P = \sum_{b=1}^{j} Y^{(b)}$ and $Q = \sum_{a=1}^{i} X^{(a)}$;
- if $\max(i,j) = j$, then define $P = \sum_{a=1}^{i} X^{(a)}$ and $Q = \sum_{b=1}^{j} Y^{(b)}$;

In both cases, application of Lemma 10 that $\max(P,Q) \geq_{\mathrm{st}} Q$ yields $C_{ij}^{y} \geq \Pr\left(\sum_{a=1}^{\max(i,j)} Y^{(a)} > y\right)$. Since Y and Z are i.i.d., we obtain $C_{ij}^{y} \geq D_{ij}^{y}$ for all $y \geq 0$ and for all i, j. This concludes the proof. □

Now we have the tools to show that M_n stochastically dominates R_n, as described in the following lemma.

Lemma 11. *For all $n \geq 0$, we have*

$$M_n \geq_{\mathrm{st}} R_n$$

where $M_n = \max(T_n^{(1)}, T_n^{(2)})$ as defined in eq. (B2).

Proof. We use induction on n. The base case $n = 0$ is an equality by definition of R_0. Now assume the statement from the lemma holds for $n = m$. We will show it also holds for $n = m + 1$. First, we expand the definition of T_{m+1}:

$$T_{m+1} = \sum_{k=1}^{K} \max(T_m^{(1)}, T_m^{(2)})$$

Now apply the induction hypothesis:

$$T_{m+1} \geq_{\mathrm{st}} \sum_{k=1}^{K} R_m^{(k)}.$$

Using Lemma 9 we obtain

$$\max(T_{m+1}^{(1)}, T_{m+1}^{(2)}) \geq_{\mathrm{st}} \max\left(\sum_{j=1}^{K^{(1)}} R_m^{(i)}, \sum_{j=1}^{K^{(2)}} R_m^{(j)} \right).$$

Applying Corollary 1 to the previous equation yields

$$\max(T_{m+1}^{(1)}, T_{m+1}^{(2)}) \geq_{\mathrm{st}} \sum_{k=1}^{\max(K^{(1)}, K^{(2)})} R_m^{(k)}.$$

The left-hand side of the previous equation equals M_{m+1} by definition, while its right-hand side is R_{m+1}, again by definition. This concludes the proof. □

The third step is to derive the lower bound on the mean delivery time from Prop. 3. This follows directly from Lemma 11, as expressed in the following corollary.

Corollary 2. (Lower bound from Prop. 3) *For $n > 0$, it holds that*

$$E[T_n] \geq \frac{1}{p_{\text{swap}}} \cdot \left(\frac{3 - 2p_{\text{swap}}}{p_{\text{swap}}(2 - p_{\text{swap}})} \right)^{n-1} \cdot \nu_0$$

where ν_0 is given in eq. (B4) *or eq.* (B5)*, depending on whether elementary-link generation is modelled following a geometric or exponential distribution, respectively.*

Proof. By Wald's Lemma [58], it follows from the definition of T_n for $n > 0$ that $E[T_n] = E[K] \cdot E[M_{n-1}] = \frac{1}{p_{\text{swap}}} \cdot E[M_{n-1}]$. A lower bound on $E[M_n]$ follows from Lemma 1 and Lemma 11, resulting into

$$E[T_n] = \frac{1}{p_{\text{swap}}} \cdot E[M_{n-1}] \geq \frac{1}{p_{\text{swap}}} \cdot E[R_{n-1}].$$

The proof finishes by substituting $E[R_{n-1}]$ by the right-hand side of eq. (B3). □

2. Proof of lower bound in Proposition 3(b)

Here, we provide the expression for m_{lower} in Prop. 3(c), which is a lower bound to the mean of the delivery time after both input links are ready, but before the entanglement swap. Formally, m_{lower} is a lower bound to the mean of M_{n-1} from eq. (B2). Such a bound follows directly from Lemma 11 by the fact that $X \geq_{\text{st}} Y$ implies $E[X] \geq E[Y]$ (see Lemma 1):

$$m_{\text{lower}} = E\left[R_{n-1}\right]$$

and $E[R_{n-1}]$ is given in eq. (B3).