We develop techniques to analyse the statistics of completion times of non-deterministic elements in quantum entanglement generation, and how they affect the overall performance as measured by the secret key rate. By considering such processes as Markov chains, we show how to obtain exact expressions for the probability distributions over the number of errors that a network acquires, as well as the distribution of entanglement establishment times. We show how results from complex analysis can be used to analyse Markov matrices to extract information with a lower computational complexity than previous methods. We apply these techniques to the Innsbruck quantum repeater protocol, and find that consideration of the effect of statistical fluctuations tightens bounds on the secret key rate by 3 orders of magnitude. We also use the theory of order statistics to derive tighter bounds on the minimum quantum memory lifetimes that are required in order to communicate securely.

I. INTRODUCTION

The ability to construct large-scale quantum networks between two or more parties is a necessary precursor to the general deployment of entanglement-based quantum key distribution as a ubiquitous alternative to classical encryption [1, 2], as well as the creation of measurement-based quantum computers [3]. Implementations of such networks would range from Bell states for point-to-point communication over large distances [4, 5], to highly connected cluster states [6] and a complete distributed quantum Internet [7]. Many theoretical proposals have been put forward for different schemes to implement these tasks, and in general the construction of these quantum networks requires the use of probabilistic elements. For example, many probabilistic methods for the generation of entanglement between nodes of a network have been proposed [8–11], as well as many high-level schemes that take advantage of such methods, such as entanglement-based quantum repeaters [12–17] and entanglement distillation [18, 19]. Probabilistic methods are also used in the implementation of non-linear unitary operations on optical states, such as those used in linear optical quantum computation [20, 21] and code-based repeaters [22, 23], as well as schemes for making measurements of states in a way that is protected against particle loss [24]. The presence of such probabilistic components means that a complex composite protocol will likely take many attempts before completing its task. When a single element fails, this could result in the entire process, or a subsection of it, needing to be restarted. It may also result in waiting errors. This is where one part of the protocol finishes, but accumulates errors while waiting for another part to complete.

Typically, in many analyses of quantum network systems, the full depth of statistical information that may be gleaned from the full probability distributions over completion times or error distributions is neglected in favor of a simpler analysis, such as analysing the average values. However, this can result in too limited a characterization of the protocol, and one that may miss essential features. For example, a situation that is commonly considered in the context of quantum communication is the time taken to generate a set of entangled states between Alice and Bob, which may be distilled in order to generate a smaller number of higher-fidelity pairs. If we consider that all pairs connect after some average time, \( t \), then the secret key rate will scale linearly with the number of states that we are trying to connect in parallel, and inversely proportional with \( t \). In reality, not all pairs of entangled states will establish at the same time. However, if we intend to use all of them for distillation, then the pairs that establish first will have to be stored on quantum memories, and the fidelities of these states will decay while they wait for the other pairs to complete. It will therefore not necessarily be advantageous to have a greater number of pairs try to establish their entanglement in parallel. A good understanding of the distribution of times taken by a protocol and the error probabilities is thus essential for any analysis of a protocol.

We begin with some general methods that may be used for the analysis of probabilistic processes using Markov chain analysis. Markov chains have recently been applied to quantum networks by Shchukin, Schmidt and van Loock [26]. We build upon these techniques in order to include errors in a natural way, as well as introducing new analytic techniques to greatly reduce the computational burden that comes with any deep analysis of Markov chains. In Section II, we explain how one may construct Markov matrices for probabilistic processes, and how the matrices for larger compound processes can be constructed from the matrices of smaller processes. We show also how we can find \( \{p_t | t \in \mathbb{N}\} \) from such matrices, where \( p_t \) is the probability that the process will complete at time \( t \). In Section III, we show how one may find the probability-generating function (PGF) from the Markov matrix. We then show how one may solve the PGF to find the completion time distribution such that the computational complexity of finding \( p_t \) is decreased.
by a factor of \( n \) compared to using the matrix alone, where \( n \) is the dimension of the matrix.

In Section IV we show how to calculate the probability distribution for the number of times that a given event in a process occurs. This rather general method may be used to calculate the distribution of the number of errors that will accumulate in the running of a process, both on average and conditioned on the completion time.

In Section IV we examine a modification of the Innsbruck protocol for distillation-based quantum repeaters [27], where the available quantum memories at a repeater station are bunched. By this, we mean we separate the available pairs of quantum memories between each pair of repeater stations into bunches of fixed size which are then distilled once all entanglement connections within the bunch have completed. We apply the techniques developed here to estimate the best values for the sizes of these bunches. This allows for a richer characterization of the secret key rates reachable by a protocol than may be learned from an analysis that does not account for the statistical factors that are captured by the Markov chain formulation. Finally, in Section IV we consider a simplification of our statistical analysis of the Innsbruck protocol. By considering bounds on the order statistics of completion times of certain elements within the protocol, we derive bounds on the secret key rates. This allows us to identify minimum experimental parameters that must be reached in order to securely communicate over a repeater network of many sections.

II. MARKOV CHAINS

Let \( \mathcal{P} \) be some process that may be decomposed into events taking place across a series of discrete time-steps. This process may be summarised by a directed graph, \( G_\mathcal{P} \), which is a flowchart showing possible paths of progression. Each node represents a unique state that the process may be in at any one time. The edges leading away from each node are the possible events (with the traversal of an edge being considered to take one time-step), with the weight of each edge representing the probability that that step will be taken. Each graph must include at least one terminating node (with no edges leaving it) representing the termination of the process. For example, if \( \mathcal{P} \) is the protocol of establishing entanglement between a single pair of quantum memories by a probabilistic process that succeeds with probability \( p \), then \( G_\mathcal{P} \) is given by Fig. 1 (see [28]). Since \( \mathcal{P} \) is probabilistic, the time that it takes to complete is represented by a random variable, \( T \), that takes on value \( t \) with probability \( p_t \).

From here, we can form the square Markov matrix for the process, \( M_\mathcal{P} \), which is the adjacency matrix of \( G_\mathcal{P} \) [29]. That is to say, \( M_\mathcal{P} \) is the weight of the edge of \( G_\mathcal{P} \) leading from node \( j \) to node \( i \) for some fixed labeling of \( G_\mathcal{P} \). This immediately gives us an operational method to find \( \{ p_t \} \); if we let \( I_\mathcal{P} \) be the set of indices for terminating nodes, then we may say that

\[
M_\mathcal{P} = M_{\mathcal{P}_1} \otimes M_{\mathcal{P}_2}.
\]

We may instead wish to wait until both \( \mathcal{P}_1 \) and \( \mathcal{P}_2 \) have completed before considering \( \mathcal{P}_3 \) to have completed. In this case we should add an element to the matrix for each subprocess that keeps the system on that terminating node until the other subprocess has completed. The composite matrix is therefore

\[
M_{\mathcal{P}_3} = [M_{\mathcal{P}_1} + \text{diag}(I_{\mathcal{P}_1})] \otimes [M_{\mathcal{P}_2} + \text{diag}(I_{\mathcal{P}_2})] - \text{diag}(I_{\mathcal{P}_1} \otimes I_{\mathcal{P}_2}).
\]

where \( \text{diag}(I_{\mathcal{P}_i}) \) is a matrix with the elements of \( I \) on the diagonal, and with zeros elsewhere.

Suppose that instead we consider \( \mathcal{P}_3 \) to consist of \( \mathcal{P}_1 \) followed by \( \mathcal{P}_2 \). When we reach the terminating nodes of \( \mathcal{P}_1 \), the next time-step will have us arrive at the first node of \( \mathcal{P}_2 \). Then

\[
[M_{\mathcal{P}_3}]_{i,j} = [M_{\mathcal{P}_1} \oplus M_{\mathcal{P}_2}]_{i,j} + \sum_{k \in I_{\mathcal{P}_1}} \delta_{i,k} \delta_{j,n_{\mathcal{P}_1}+1}
\]

where \( n_{\mathcal{P}_i} \) is the number of nodes in \( G_{\mathcal{P}_i} \) or the number of rows or columns in \( M_{\mathcal{P}_i} \).

FIG. 1: Graph for simple probabilistic entanglement generation. Each node represents a state that the system may be in at any one time. Transitions between nodes are considered to all take the same length of time, and occur with probabilities indicated by the weight of the edge. Terminating node shown in black.
It may also be the case that different parts of a process take different lengths of time, instead of the above construction which assumes that each event takes a single time-step. Suppose that within some process, $\mathcal{P}$, we have some events (edges on $G_\mathcal{P}$) that take some time $k_1$ and some that take $k_2$, where $k_2 \geq k_1$. We can decompose $M_\mathcal{P}$ as $M_{\mathcal{P},k_1} + M_{\mathcal{P},k_2}$, such that all elements in $M_{\mathcal{P},k_1}$ represent events that take $k_1$, and similar for $M_{\mathcal{P},k_2}$. From this we can create a new process matrix $M'_\mathcal{P}$ which properly accounts for the fact that events in subprocess $P_1$ can be done many times for each time that $P_2$ can be done. This is given by

$$
[M'_\mathcal{P}]_{i,j} = \sum_k \left\{ [M_{\mathcal{P},k_1}]_{i,k} + \left( 1 - \sum_i [M_{\mathcal{P},k_2}]_{i,k} \right) \delta_{i,k} \right\} [M_{\mathcal{P},k_2}]_{k,j} \tag{5}
$$

As such, $p_t$ calculated from $M'_\mathcal{P}$ will represent the probability that the process completes after $t$ applications of $P_2$ and $tk_2/k_1$ applications of $P_1$. It should be noted that the modification of process matrices to account for timing differences should be done before creating composite matrices by tensor products.

Using Markov matrices along with Eq. (1) is a simple way to calculate the completion times of a process, although it is not necessarily the most efficient. Multiplying $M_\mathcal{P}$ by itself takes $n_\mathcal{P}^2$ elementary multiplications. Given some algorithm for calculating exponentials that has a number of operations that scales asymptotically as $f_{\text{exp}}(t)$ for the calculation of $k^t$ for some constant $k$, we find that the calculation of $p_t$ scales asymptotically as

$$
\mathcal{O}(p_t \text{ by matrix mult}) = \mathcal{O}(n_\mathcal{P}^2 f_{\text{exp}}(t)) \tag{6}
$$

In the next section we derive a method by which this may be reduced by a factor of $n_\mathcal{P}$.

III. PROBABILITY GENERATING FUNCTIONS

In this section we show how an approach based on probability generating functions (PGFs) and complex analysis can lead to formulas for $p_t$ that are faster to compute than the matrix multiplications of Eq. (1).

The probability generating function of a distribution $\{p_t\}$ corresponding to the completion times for a process $\mathcal{P}$ is defined as the polynomial

$$
f_\mathcal{P}(z) = p_0 + p_1 z + p_2 z^2 + p_3 z^3 + \cdots \tag{7}
$$

where $z$ is a complex variable, and $p_t$ are constants to be determined based on $\mathcal{P}$. Given the PGF associated with some process, the elements $p_t$ may be found by calculating the coefficients of the various terms by finding the derivatives:

$$
p_t = \left. \frac{1}{t!} \frac{d^t f_\mathcal{P}(z)}{dz^t} \right|_{z=0}. \tag{8}
$$

In order to write down the PGF, it may seem like we need to already know all of $p_t$. However, we can calculate $f_\mathcal{P}(z)$ directly from $G_\mathcal{P}$. Consider a node in $G_\mathcal{P}$, $x$, with one edge leading to node $y$ with probability 1. Let $f_\mathcal{P}^{(x)}(z)$ be the PGF for the system when we start at node $x$. Since the system will take exactly one time-step longer to complete when we start at $x$ than when we start at $y$, we can say that $f_\mathcal{P}^{(x)}(z) = z f_\mathcal{P}^{(y)}(z)$. Now suppose that $x$ has two edges leading away from it to nodes $y_1$ and $y_2$ with probabilities $p(y_1)$ and $p(y_2)$ respectively. Then, $f_\mathcal{P}^{(x)}(z) = z p(y_1) f_\mathcal{P}^{(y_1)}(z) + z p(y_2) f_\mathcal{P}^{(y_2)}(z)$. By extension, we may say that

$$
f_\mathcal{P}^{(i)}(z) = \begin{cases} \sum_j [M_\mathcal{P}]_{i,j} z f_\mathcal{P}^{(j)}(z) & \text{if } i \not\in I_\mathcal{P}, \\
1 & \text{if } i \in I_\mathcal{P}, \tag{9}
\end{cases}
$$

where the sum runs over all columns in the matrix, which is an eigenvalue equation. The PGF of the process as a whole $[f_\mathcal{P}(z)]$ may be identified with the PGF of the initial node $[f_\mathcal{P}^{(1)}(z)]$. In particular, $f_\mathcal{P}(z)$ is the first element of the eigenvector of $M_\mathcal{P}(z)$ with eigenvalue 1, normalised such that the $k^{th}$ element is 1 for any $k \in I_\mathcal{P}$, where

$$
\tilde{M}_\mathcal{P}(z) = z M_\mathcal{P}^T + \text{diag}(I_\mathcal{P}). \tag{10}
$$

However, a problem may arise in the process of finding the set of eigenvectors. We wish to retain $z$ as an open variable in the PGF, which means that many of the fast methods for finding eigenvalues of matrices cannot be used, since they rely on numerical methods. In order to find an eigenvector of a completely general matrix, $M$, we need to be able to solve the characteristic equation $|M - \lambda I| = 0$. This involves exactly solving a polynomial of order $n_\mathcal{P}$, which cannot in general be done for $n_\mathcal{P} \geq 5$. Instead, we use the fact that the eigenvalue is 1, so that $M f_\mathcal{P} = f_\mathcal{P}$, where $f_\mathcal{P}$ is the vector with $i^{th}$ element equal to $f_\mathcal{P}^{(i)}$, and say that

$$
f_\mathcal{P}(z) = [f_\mathcal{P}]_1 / [f_\mathcal{P}]_k
$$

$$
f_\mathcal{P} = \text{Null} \left[ \tilde{M}_\mathcal{P}(z) - 1 \right], \tag{11}
$$

for any $k \in I_\mathcal{P}$. Note that we have used a slight abuse of notation and specified that $f_\mathcal{P}$ is equal to the null space itself and not a particular vector in the null space. This is because the null space has a dimension of 1. We can see this by the fact that, if $M_\mathcal{P}$ is a Markov matrix, then
\[ \frac{1}{i!} \frac{d^i \mathcal{P}(z)}{dz^i} \bigg|_{z=a} = \frac{1}{2\pi i} \oint_{\partial \mathcal{S}} \frac{\mathcal{P}(z)}{(z-a)^{i+1}} dz, \] (12)

where \( \partial \mathcal{S} \) is the boundary of \( \mathcal{S} \); a compact subset of \( \mathbb{C} \) on which \( \mathcal{P}(z) \) is analytic.

Let \( a = 0 \) and \( g_i(z) = \mathcal{P}(z)/z^{i+1} \). We will evaluate the integral of \( g_i(z) \) on a circle centered on \( z = 0 \). If the contour encloses no poles except the one at 0 due to the \( z^{-(i+1)} \) term, then this is equivalent to finding the residue of the pole of \( g_i(z) \) at 0. Suppose that \( \mathcal{P}(z) \) scales as \( O(z^0) \) as \( |z| \to \infty \). Then the integral of \( g_i(z) \) on a circular path of radius \( R \) will tend to 0 as \( R \to \infty \) for all \( t > t_0 \) (since \( dz = |z|d\theta \)). However, by Cauchy’s residue theorem, this integral must also be equal to the sum of all residues of \( g_i(z) \) in \( \mathcal{S} \). This includes the pole at 0, which we get from the \( z^{-(i+1)} \) term, and the poles elsewhere in the complex plane, which are the poles of \( \mathcal{P}(z) \). Therefore the sum of the residues of all poles must be equal to 0 for \( t > t_0 \). The residue at \( z = 0 \) cannot be calculated directly since it is a non-simple pole, but we can calculate it indirectly since we know it must be equal to the negative of the sum of the residues of the other poles, which are in general simple. We therefore arrive at the main result of this section:

\[ p_t = -\sum_i \text{Res} \left[ \frac{\mathcal{P}(z)}{z^{i+1}}, z_i \right] \in \mathcal{P}(f_P), \] (13)

where \( \mathcal{P}(f) \) is the set of singularities of \( \mathcal{P}(z) \).

As a corollary, we may use this method to easily find expectation values for completion times of such processes. Consider that

\[ \langle T \rangle = \sum_t t \ p_t, \] (14)

If we use the fact that

\[ \text{Res} \left[ \frac{\mathcal{P}(z)}{z^{i+1}}, z_i \right] = \frac{\text{Res} \left[ \frac{\mathcal{P}(z)}{z}, z_i \right]}{z_i^{i+1}}, \] (15)

since \( f_P \) has no pole at 0, we can write Eq. (14) as

\[ \langle T \rangle = -\sum_i \text{Res} \left[ \frac{\mathcal{P}(z)}{z}, z_i \right] \sum_t \frac{t}{z_i^{i+1}}, \] (16)

where the \( z_i \) sum is implicitly over the poles of \( f_P(z) \). We can use the identity \( \sum_{n=0}^{\infty} nx^{n-1} = (1-x)^{-2} \) to find the sum over the \( t \)-dependent terms, giving

\[ \langle T \rangle = -\sum_i \text{Res} \left[ \frac{\mathcal{P}(z)}{z}, z_i \right] \frac{1}{(1-z_i)^2}, \] (17)

Similarly, we can find the probability of the process completing by \( t \), and the variance of the completion times:

\[ p(T \leq t) = \sum_i \frac{1 - z_i^{t-1}}{1 - z_i} \text{Res} \left[ \frac{\mathcal{P}(z)}{z}, z_i \right], \] (18)

\[ \text{Var}(T) = \sum_i \frac{1 + z_i}{(1 - z_i)^3} \text{Res} \left[ \frac{\mathcal{P}(z)}{z}, z_i \right] - \langle T \rangle^2, \] (19)

We may note now that, if \( f_P(z) \) is built up constructively, as in Eq. (6), each non-terminating node contributes a single factor of \( z \) to the PGF. This means that \( f_P(z)z \) must be of order \( n_P - |I_P|^2 \) at most, and so have no more than \( n_P - |I_P|^2 \) poles. When calculating \( p_t \) by Eq. (13), the residues of \( f_P(z)/z \) are not \( t \)-dependent. Therefore, when we vary \( t \), we simply need to calculate \( z_i^t \) for each \( z_i \in \mathcal{P}(f_P) \). Given again some algorithm for calculating exponentials \( k^t \) in \( O(f_{\text{exp}}(t)) \) operations, we have that calculating \( p_t \) now scales asymptotically as

\[ O(p_t \text{ by Cauchy}) = O\left( \left| n_P - |I_P|^2 \right| f_{\text{exp}}(t) \right), \] (20)

which represents an improvement of a factor of \( n_P \) over the matrix multiplication method.

### IV. ERROR DISTRIBUTIONS

In any process, there will be events that have some probability to cause an error. For example, if an event represents a state being stored on a quantum memory, then in each time-step there is some non-zero probability that the memory fails and the information stored on it is lost. When carrying out the process, we wish to know \( p(k|t) \); the probability that we will pass such an edge \( k \) times, given an overall process completion time of \( t \). This implicitly assumes that such a process is heralded. That is, we always know what stage of \( \mathcal{P} \) we are at, and so
can count the number of occurrences of an error-carrying event. A non-heralded process would be one where we have a description of $G_P$, but we do not know how close we are to completion at any time, but instead are simply informed when the process completes. If each occurrence of an error-carrying event has a probability $\epsilon$ to cause an error, then the overall probability that an error will have occurred is given by

$$p(\text{error heralded}) = 1 - (1 - \epsilon)^k,$$

$$p(\text{error non-heralded}) = \sum_{k=0}^{\infty} p(k|t) [1 - (1 - \epsilon)^k]. \quad (21)$$

In order to include this in our analysis, we must first identify which events (edges in $G_P$) may cause an error. Then, for each event in question between edges $j$ and $i$, we multiply $[M_P]_{i,j}$ by an open complex variable, $w$, which will call the counting variable.

Now note that the value for $p_t$, calculated either by Eq. (1) or Eq. (12), may be seen as a sum of the probabilities of the different sequences of events by which the process may be completed in time $t$. When one term in $M_P$ is an open variable, $p_t$ will be expressed as a finite polynomial in $w$, which we will denote $p_t(w)$. For such sequence of events that includes $k$ passes of an error-carrying edge and occurs with probability $p_a$, $p_t(w)$ will include a term equal to $p_a w^k$. The full expression for $p_t(w)$ will then be of the form

$$p_t(w) = p_t(1) \sum_{k=0}^{O(p_t(w))} p(k|t) w^k. \quad (22)$$

where $O(p_t(w))$ is the order of $p_t(w)$. Thus by finding the coefficients of this polynomial, we can find the error distributions. This polynomial is finite, with all terms involving $w$ to a non-negative power. Therefore it has no poles, so we cannot use the methods of Section 11. Instead we can extract the coefficients by way of a (fast) Fourier transform, which, unlike the complex analysis method, can be done numerically. To do this, we first should identify some number $N_P(t)$ such that $O(p_t(w)) \leq N_P(t) \leq 2O(p_t(w))$, where the latter inequality is to avoid aliasing effects 32. We then evaluate $p_t(w)$ at $N_P(t)$ equally spaced complex points, given by $\{p_t(e^{i2\pi k/N_P(t)})|k = 1, \ldots, N_P(t)\}$. The discrete Fourier transform of these evaluated points reveals the first $N_P(t)$ coefficients of $p_t(w)$ (where all greater coefficients are 0). Applications of this Fourier method for extracting coefficients to more general analytic functions are described in 33.

This construction may also be used to account for different kinds of errors, by multiplying matrix elements by different complex variables, $w_1, w_2, w_3, \ldots$, and performing a multi-dimensional Fourier transform on $p_t(w_1, w_2, w_3, \cdots)$ to determine $p(k_1, k_2, k_3, \cdots | t)$.

\begin{figure}[h]
\centering
\includegraphics[width=0.7\textwidth]{fig2}
\caption{Innsbruck quantum repeater scheme with 2 sections. This shows the entanglement generation scheme for one “bunch,” (defined in the last paragraph of Section 1) where $q_0 = 4$. Part 1 shows a situation where some pairs of memories have established entanglement, where $A$ is Alice, $R$ is Richard and $B$ is Bob. Circles show quantum memories dotted lines show no entanglement and solid lines are successful entanglement connections. In parts 2–3, the right-hand section has completed, and distilled down to $q_1 = 2$ pairs. The thickness of the lines between memories shows the fidelity of that entangled state. In parts 4–5, the left-hand section is also distilled and these states are connected to the right-hand pairs by entanglement swapping. In part 6 the two pairs between Alice and Bob are distilled again to form a single high-fidelity pair.}
\end{figure}

From this we can also read off the average error rate for a process completing by time $t$, i.e. the non-heralded error. Suppose we have only a single type of error. From Eqs. 21 and 22, we can say that

$$p(\text{error non-heralded}) = \frac{1 - p_t(w = 1 - \epsilon)}{p_t(1)} \quad (23)$$

The case for multiple types of error follows as a simple extension of this.

V. INNSBRUCK PROTOCOL ANALYSIS

In this section we will be considering the repeater protocol of the Innsbruck group 27. In the standard implementation of this protocol, there are $Q_0$ pairs of quantum memories between each adjacent pair of repeater stations. These are all connected in parallel, and then
distilled to make \( Q_1 \leq Q_2/2 \) pairs. By entanglement swapping, these are then connected with adjacent pairs to form entanglement over twice the length, and distilled again to form \( Q_2 \leq Q_0/4 \) pairs in parallel between each section, and so on. Previous analyses of this protocol have either assumed that the entanglement connection can be done almost deterministically, or considered that entanglements establish after some average time. So if each attempt to establish entanglement between two stations has a probability to succeed of \( p \), then a simplified approach to understanding the system and estimating the key rate would be to assume that all pairs establish entanglement after \( 1/p \) attempts.

Considerable progress has been made in understanding and building upon the Innsbruck protocol, since it is one of the most promising routes to constructing long-distance quantum communication. Much of this work has focused on aspects such as the relation between the key rate and experimental imperfections \[34\], the specifics of how to implement the system with atomic ensembles \[35\], understanding and improving the robustness against channel noise \[15, 36\] or side-channel attacks \[37–40\]. However, one important aspect is often overlooked, namely the statistical factor of waiting times arising from probabilistic completion times of different elements. We now show that this has severe implications for the performance of the protocol.

In our analysis, we let the \( Q_0 \) parallel pairs of memories between each pair of repeater stations be divided into bunches of \( q_0 \) pairs. When all pairs within such a bunch have completed, then they are distilled to \( q_1 \leq q_0/2 \) pairs. This is an inequality, since distillation (as described more fully in Appendix \[B\]) is a probabilistic process. This is shown in Fig. 2. Here, there is a trade-off inherent in the size of \( q_0 \). When \( q_0 \) is small, the bunch will complete quickly on average. This means that the first entangled pair to complete will not have to wait long before the last one completes, and so is less likely to accrue memory errors. However, one then has fewer options for distilling a high fidelity state. Given a large set of states, we can instead find a better optimal strategy for combining states under a distillation protocol to result in a higher final secret key rate, at the expense of longer waiting times.

A. Constructing the matrix

Here we consider at first a repeater consisting of two sections, separated by a distance \( L \). Alice tried to establish entanglement between herself and Richard (a repeater station), and Richard between himself and Bob. The Markov graph for the establishment of a single Bell pair is shown in Fig. 1. Let its Markov matrix be \( M_{\text{Bell}} \). We shall consider one time-step in this process to be \( 2L/c \), where the factor of 2 arises since the receiving party needs to send a classical signal back to the sending party to confirm whether the previous photon was received or not.

We now use Eq. \[2\] to construct the matrix for \( q_0 \) pairs connecting in parallel between a pair of repeater stations. We want to include a complex counting variable, \( w_0 \) that counts how many time-steps a given quantum memory has to wait before the others finish. However, we should note that we include this only on a single factor of the matrix for the section, \( M_{\text{sect}} \), to avoid multi-counting errors. The counter \( w_0 \) therefore counts how many errors accumulate on a particular entanglement link. By symmetry, we can say that this error distribution is equal across all such entanglement links. Therefore,

\[
M_{\text{sect}} = [M_{\text{Bell}} + w_0 \text{ diag}(1_{\text{Bell}}) \otimes [M_{\text{Bell}} + \text{ diag}(1_{\text{Bell}})]^{\otimes q - 1}. \]

where

\[
M_{\text{Bell}} = \begin{bmatrix}
1 - p & 0 \\
p & 0
\end{bmatrix}, \quad 1_{\text{Bell}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.
\]

where \( p \) is the probability of entanglement being established in any particular attempt. In Appendix A, we describe how this may be reduced in dimension by symmetry arguments.

Once all pairs between two stations have established their entanglement, we want to perform a distillation on these. These are matched up into \( \lfloor q_0/2 \rfloor \) pairs, which are then distilled using the DEJMP protocol \[18\]. A DEJMP distillation between two noisy Bell pairs has some non-unity chance of success, which depends on the fidelities of the states involved. However if the success probabilities were fidelity-dependent, that would mean including terms in the matrix which depend on the time taken for the process to reach that event, a modification which would move us outside the realm of Markovian dynamics. Therefore we will choose some minimum distillation success probability, \( \lambda \), corresponding to the success probability two states of fidelity \( F_{\text{min}} \) being distilled with each other. We will later exclude any runs of the process that would have used states of fidelity less than \( F_{\text{min}} \), as explained in Appendix B. In this way, any choice of \( \lambda \) will give us a lower bound on the secret key rate, and we may freely maximise over choices of \( \lambda \). We therefore add a row and column to \( M_{\text{sect}} \) to form \( M_{\text{dist}} \). The new column has an element representing distillation success, with a probability of \( 1 - (1 - \lambda)^{\lfloor q_0/2 \rfloor} \). The “failure” event (of all distillations failing) resets the process of creating entanglement on that section. Therefore, if \( M_{\text{sect}} \) in Eq. \[24\] takes the form

\[
M_{\text{sect}} = \begin{bmatrix}
M'_{\text{sect}} & 0 \\
\alpha^T & 0
\end{bmatrix},
\]

for some matrix \( M'_{\text{sect}} \) and vector \( \alpha \), then \( M_{\text{dist}} \) is given by
where the vector $b = [(1 - \lambda)|q_0/2|, 0, 0, \ldots, 0]^T$, and $s = 1 - (1 - \lambda)|q_0/2|$.

Finally, we construct the matrix for the entire system, $M_{\text{full}}$, in a similar way to Eq. (24) by considering two copies $M_{\text{dist}}$. We again include a complex counting variable to account for all memories on one section needing to wait until the other side has been connected and distilled. For this we make sure to use a different counting variable, $w_1$, so we can keep track of the distribution of errors that occur before and after the first round of distillation.

B. Analysis

We can now analyse $\{p(k_0, k_1 | t)\}$, where $k_0$ and $k_1$ are the number of passes of edges weighted by $w_0$ and $w_1$ respectively. By doing this for a fixed $q_0$, we can find a distillation strategy that gives the maximum possible achievable secret key rate for a given completion time, $t$, averaged over the error distribution (explained in detail in Appendix B), which we shall call $K(t|q_0, p, \epsilon_W)$. Since the key rate of a protocol goes inversely with the time taken to establish a raw bit, and linearly with the number of parallel “bunches” of states that are used, $Q_0/q_0$, we will use the normalised average key rate as a function of $q_0$ as our main figure of merit for the system analysis:

$$K(q_0, p, \epsilon_W) = \frac{1}{q_0} \sum_{t=1}^{\infty} \frac{p_t}{t} K(t|q_0, p, \epsilon_W) \quad (28)$$

For the error probability per time period, $\epsilon_W$, we assume the probability for a quantum memory to not undergo an error decays exponentially with time as $\epsilon_W = 1 - \exp(4L/c\tau)$, where the extra factor of 2 is due to the fact that each entangled pair involves 2 memories. Here, $\tau$ is the memory lifetime. In Fig. 3 we show the a few examples of calculated normalised key rates for different values of $q_0$. In order to compare the insight gained from this method to the estimations that might result from a less nuanced analysis, we have also shown the simplified secret key rate. Here, we consider only the average connection time of an entanglement link. That is to say, we assume that all links wait for a time $1/p$, and then connect deterministically. For a fair comparison, we have retained the same maximization over distillation strategies that is outlined in Appendix B. The simplified secret key rate may therefore be written as

$$K_{\text{simp}}(q_0, p, \epsilon_W) = \frac{1}{q_0} \sum_{t=1}^{\infty} \frac{p_t}{t + 1/p} K(t|q_0, 1, \epsilon_W) \quad (29)$$

When we consider Fig. 3 we can see that an estimation of the secret key rate that considers only the average completion time severely underestimates the performance of the protocol. This is particularly striking when we note that no errors accumulate in the simplified analysis, due to the fact that no elements are left waiting while others complete. In particular, the difference between the statistical and simple key rates in the $\epsilon_W = 1\%$ case reaches 3.8 orders of magnitude, which could mean the difference between communicating in kilobits and megabits per second.

We may also note that, for low values of $\epsilon_W$, the key rate rises with increased bunch size. This means that increasing the pool of states available to be distilled has a greater-than-linear effect on the key rate, highlighting the power and importance of distillation to quantum technologies.

VI. ANALYTIC KEY RATE BOUNDS FOR THE INNSBRUCK PROTOCOL

The techniques presented above allow for a thorough investigation of the contribution of statistical factors arising from non-deterministic protocol elements towards the secret key rate of a general quantum communication protocol. While this has been presented in-depth for a two-section repeater, practical systems will often demand the application of a series of many repeaters. The current limit for repeater-unassisted quantum communication is on the order of a hundred kilometers. If we therefore wish to securely communicate on an intercontinental scale, we require a method of analysis that can scale up to dozens of repeater sections. This presents a limitation in our protocol: while the dependence of $n_p$ on $q_0$ can be made linear (Appendix A), the dependence on the number of
sections, $N_S$, remains exponential.

In the original analysis of the Innsbruck protocol, the fidelity of the final shared state was not considered to be fundamentally dependent on the number of sections. This is because the $l^{th}$ level of the protocol, which consists of taking entangled pairs over some distance $2^l L$, distilling them, and connecting with adjacent sections to form pairs over a distance $2^{l+1} L$, would produce pairs of a fidelity that did not depend on $l$.

In addition to waiting times increasing with $q_0$, there is also the issue that the classical communication time grows with the distance over which pairs are entangled. However, if statistical factors are ignored then this could be dealt with using a “blind” protocol, where distillation and attempts are assumed to have succeeded at every stage, and communication after the termination of the protocol allows for a post-selection on the attempts that succeeded. This allows for the final fidelity to be kept above the minimum level required for secure communication at the expense of a hit to the raw key bit generation rate.

Once we include the statistical waiting times in our analysis, it is no longer true that a non-zero secret key rate can be guaranteed for all length scales. In this section we examine the behaviour of a repeater network with a minimum requirement of physical qubits, which is $q_0 = 2^{N_S}$. Instead of finding best key rates, we look here for the parameters for which the secret key rate is lower bounded above zero. By finding the threshold parameters that need to be reached for the protocol to operate, we can identify concrete values for component designers to aim for, and give benchmarks by which we can compare performances.

We will consider that after every quantum operation, a Werner twirling procedure is applied to all states. This involves unitarily mapping the state to the singlet state, $|\Psi^-\rangle$, then applying a randomly chosen local unitary Pauli operation identically each part of the entangled pair. This maps all states to Werner states of the same fidelity (see Appendix B). By doing this, we can simply consider the effect of the repeater network as a recursive function on a single real variable - that of the average fidelity. Note that this operation is not actually carried out, it is simply used to repeatedly map states to the analytically simple Werner states. This may be done since applying local operations cannot increase the strength of entanglement by any measure, and so cannot increase the secret key rate.

The analysis then proceeds as follows. Level 0 of the protocol consists of all pairs within one section connecting at initial fidelity $F_{\text{init}}$. Instead of considering the full probability distribution of waiting times, we consider that all pairs wait for a number of time-steps equal to the estimated time for the last pair to connect, $k_L$, which upper bounds the waiting time for each pair. This time is equal to the expectation value for the largest order statistic from a sample of $q_0$ chosen from the distribution with cumulative distribution function $1-(1-p)^q$. By choosing $p \ll 1$, such that we may allow the distributions to be approximated by continuous functions, we may use results from [42] to bound this by

$$k_L \leq \left( \frac{q-1}{\sqrt{2q-1}} + 1 \right) \frac{1}{|\log(1-p)|} + 1. \quad (30)$$

These states are then distilled to produce states of fidelity

$$F_0 = J[D(F_{\text{init}}, \epsilon_W, k_L)], \quad (31)$$

where the functions describing the effect of the decay of quantum memories over time $k_L$ on the average fidelity and DEJMPS distillation are given respectively by

$$D(F, \epsilon_W, k_L) = \left( 1 - \epsilon_W \right)^{k_L} F + \frac{1 - (1 - \epsilon_W)^{k_L}}{4},$$

$$J(F) = \frac{10F^2 - 2F + 1}{8F^2 - 4F + 5}, \quad (32)$$

respectively. The $l^{th}$ level of the protocol consists of the following when $l \geq 1$. Within each pair of two sections, one section will complete first, and wait for a time no longer than $k_{A,l}$ for the latter to complete. We show in Appendix C that this is bounded by

$$k_{A,l} \leq 2^l \left[ \frac{H(2^{N_S-l+1})}{|\log(1-p)|} + 1 \right], \quad (33)$$

where $H[n] = \sum_{m=1}^{n} 1/m$ is the $n^{th}$ harmonic number. The average fidelity after level $l$ can then be defined recursively as

$$F_l = J(C(F_{l-1}, \tilde{F}_{l-1}, \epsilon_L)), \quad (34)$$

where $C$ gives the average fidelity after connecting two adjacent sections and twirling, where we have allowed here for local gate errors. This is given by

$$C(F_a, F_b, \epsilon_L) = D \left( \frac{1}{3} (1 - F_a)(1 - F_b) + F_a F_b, \epsilon_L, 1 \right), \quad (35)$$

where $\epsilon_L$ is the probability that an error occurs when performing the local operations involved in entanglement swapping, and $\tilde{F} = D(F_{l-1}, \epsilon_W, k_{A,l-1})$.

In Fig 4 we show the minimum quantum memory lifetimes required for a non-zero key rate as a function of $p$. From this it can be seen that the probability for an entanglement attempt to succeed is the biggest factor in affecting the ability to securely communicate. For comparison we also include the requirements for the case that does not include statistical effects, where $k_{A,l} = 2^l$ for all $l$. It can be seen that the minimum memory requirements
FIG. 4: Minimum quantum memory lifetime, 
\[ \tau = -4L/\log(1 - \epsilon_W) c \], required for a
minimum-resources quantum repeater to securely
communicate over a network of 8 sections of length
\( L = 25\text{km} \). \( F_{\text{init}} \) is the initial fidelity of entangled pairs
before distillation or connection, and \( N_S \) is the number
of sections. Solid lines indicate the cases that include
statistical factors, and dashed lines do not. Both solid
and dashed lines are ordered the same in the plots and
the legend. Blind distillation is not used in either case.

are slightly higher in the case where statistical effects are
included, but this effect decreases with the number of
sections over which we connect. For an even comparison,
we have not used blind distillation in the non-statistical
case. We see that there is a constant-factor increase in
the required lifetime of the memories. In some cases this
reaches as high as a factor 2 increase in the required life-
time of the quantum memories. The resultant bounds are
just reachable by the lifetimes of atomic ensembles, which
can have lifetimes up to 40ms [43]. However, all bounds
are well within the lifetimes of the nuclear spin states of
NV centers [44]. This implies that the main challenge
is the construction of optical elements with high transmis-
sion and detection efficiencies, whereas NV-center-based
protocols [15, 45] may be more suitable when these effi-
ciencies are low.

VII. CONCLUSION

Many of the practical quantum technologies that are
being proposed are inherently probabilistic in nature,
which leads to uncertain completion times and error dis-
tributions. We have developed techniques that allow
for thorough characterizations of such statistical distrib-
utions in both the computational and analytic directions. In terms of computational techniques, we have
used Markov chains to analyse quantum entanglement
generation. We have shown how to form composite sys-
tems from smaller elements in a constructive manner.
We have then shown how to use the Markov matrices
for such composite systems to calculate the probability
distribution over the number of errors that occur in the
running of a general protocol. This allows for a com-
plete characterization of the fidelities of the states that
are produced by a quantum protocol. As an example, we
have analysed the Innsbruck quantum repeater protocol
with a memory-error model. A thorough understanding
of the set of resultant errors has been shown to lead to
a tighter bound on the secret key rate than an analysis
based only on an averaged approach. In some cases this
resulted in tightening the bounds on key rates by over 3
orders of magnitude – a clear indication that a considera-
tion of statistical effects does not simply provide a minor
correction to performance, but instead is fundamental to
understanding the quantitative behaviour of a system.

In terms of analytic techniques, we have shown how
elements of the eigenvectors of a transformed form of the
Markov matrix correspond to the probability generating
function of the process. This has been solved for an ar-
bitrary term in the probability distribution over comple-
tion times by using results from complex residue analysis.
This was done in a way such that the number of computa-
tional operations required scales only linearly with the
number of states in the Markov process, compared to the
quadratic scaling of a more direct approach. Finally we
have shown how the theory of order statistics can put
bounds on the statistical effects on the secret key rate,
and used this to bound the minimum quantum memory
lifetimes needed to run the Innsbruck repeater protocol.

ACKNOWLEDGEMENTS

This research was made possible via the EPSRC Quan-
tum Communications Hub, Grant No. EP/M013472/1

Appendix A: Simplifying Markov matrices with high
symmetry

The Markov matrix for a single section given in
Eq. (24) is \( 2^n \)-dimensional. While this accurately de-
scribes the dynamics of the system, we can take advan-
tage of the fact that the system contains a high degree of
symmetry to reduce the size of the matrix. We can use
the fact that the probability to move between one state
and another is only dependent on how many en-
tanglements have been established in the initial and final
states, and not on the specifics of which entanglements.
We can therefore use a technique called “lumping,” where
we create a partition of the states into sets, as shown in
This is shown for the $q_0 = 3$ case below.

$$
\begin{bmatrix}
3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 3
\end{bmatrix}
$$

(A1)

This effectively distributes the counting variable amongst the states, giving it the symmetry required to lump the states together. This reduces the number of unique states in the process corresponding to one section from $2^4$ to $q + 1$.

Appendix B: Detailed description for the analysis of the Innsbruck protocol

The first step in analysing the modified Innsbruck protocol is to construct the associated Markov matrix, as described in Section 4A and Appendix A. To do this we fix $q_0$, which sets the size of the matrix, and $p$, which determines the elements of the matrix. For a fixed $t$, a joint distribution of $k_0$ and $k_1$ is then calculated.

We assume that the states that are initially created after establishment is connected are Werner states of the form

$$
\rho = \frac{4F_{\text{init}} - 1}{3} |\Phi^+\rangle \langle \Phi^+| + \frac{1 - F_{\text{init}}}{3} 1,
$$

where $F_{\text{init}}$ is the fidelity with respect to $|\Phi^+\rangle$. We then choose some $k_{\text{max}}$ that gives some largest acceptable error. Then from the marginal distribution on $k_0$ we then choose $2q - 1$ values for $k_0$ (all of which are below $k_{\text{max}}$), and we choose a final value for $k_0$ and a value for $k_1$ from the full error distribution. These transform the $2q$ states (by $q$ on each section of the repeater network) as:

$$
\rho \mapsto (1 - \epsilon_{W_0})\rho + \epsilon_{W_0} 1,
$$

where $\epsilon_{W_0}$ are defined as heralded errors as in Eq. (21), with $k = k_0$.

We now partition the set of states into a ‘left set’ and a ‘right set’, corresponding to the two different section of the network, and randomly apply a distillation to each set. To do this, we pair up the states within a set. If $q_0$ is odd, one state is randomly chosen to proceed to the next round without being distilled. The remaining pairs are distilled according to the DEJMPS protocol, which maps states on the basis $|\Phi^+\rangle, |\Psi-\rangle, |\Psi+\rangle, |\Phi-\rangle$, with diagonal coefficients $(a_1, b_1, c_1, d_1), (a_2, b_2, c_2, d_2)$ to the state with coefficients...
where the distillation succeeds with probability \( \mathcal{N} = (a_1 + b_1)(a_2 + b_2) + (c_1 + d_1)(c_2 + d_2) \).

As explained in the main text, when translating the distillation success probability to a term in the Markov matrix, we use a constant probability of distillation success, \( \lambda \). This is related to our choice of \( k_{\text{max}} \) by

\[
k_{\text{max}} = \left[ \log \left( \frac{3\sqrt{2\lambda - 1}}{4F_{\text{init}} - 1} \right) \right. \left. \frac{1}{1 - \epsilon_{W0}} \right]
\]

where \( a, b, c, d \) are random variables, with \( \epsilon_{W1} \) calculated from \( k_1 \).

For a fixed \( q_1 \), we then calculate the secret key rate as follows. We choose a random pairing of states on the left with states on the right. They are deterministically connected by applying a CNOT gate to the part of each Bell state stored in the repeater, and then measuring each in the \( X \) basis. This maps two states of diagonal coefficients \((a_1, b_1, c_1, d_1), (a_2, b_2, c_2, d_2)\) to one with coefficients

\[
\begin{bmatrix}
  a_1 a_2 + b_1 b_2 + c_1 c_2 + d_1 d_2 \\
  a_1 b_2 + a_2 b_1 + c_1 d_2 + c_2 d_1 \\
  a_1 c_2 + a_2 c_1 + b_1 d_2 + b_2 d_1 \\
  a_1 d_2 + a_2 d_1 + b_1 c_2 + b_2 c_1
\end{bmatrix}.
\]

These final states may then be distilled again. We optimize over combinations of distillation pairings to produce \( q_2 \leq q_1 \) final pairs, in order to maximise the secret key rate, given by

\[
K(t|q_0, q_1, p, \epsilon_W) = \sum_{i=1}^{q_2} 1 - 2h_2(\epsilon_i),
\]

where \( h_2(x) = -x \log_2 x + (1 - x) \log_2 (1 - x) \) is the binary entropy function, and \( \epsilon_i \) is the bit error of the \( i \)th entangled pair, averaged between measuring in the \( Z \) basis and the \( X \) basis.

We must finally multiply \( p_t \) by the probability that none of the states involved in completing the process were of a fidelity less than \( F_{\text{min}} \). As such, we make the transformation

\[
p_t \mapsto p_t \times \left( \frac{1}{p_t} \sum_{k=k_{\text{max}}}^{t} p(k|t) \right)^{q_0}.
\]

This key rate is optimized over distillation strategies (both before and after entanglement-swapping) and entanglement-swapping pairing choices, and averaged over values of \( q_1 \) and selections of sets of \( k_0, k_1 \) from the distribution to get \( K(t|q_0, p, \epsilon_W) \) which is used to get from \( K(q_0, p, \epsilon_W) \) by Eq. (28).

Appendix C: Analytic calculation of average waiting times

The distance over which communication has to occur at level \( l \) scales with \( 2^l \). Given two sections of a repeater, there is some number of time-steps \( k_{2\text{sec}} \) between the first completing and the second. After the second section completes, there must be one more round of classical communication to indicate this fact. Therefore \( k_{A,l} = 2^l (k_{2\text{sec}} + 1) \).

We wish to calculate \( k_{2\text{sec}} \), which is given by \( E|y - x| \), where \( x \) and \( y \) are two times drawn from the distribution, \( f(t) = d_t C(t) \), where the cumulative distribution function is given by \( C(t) = [1 - (1 - p)^l]^y \). Approximating these as continuous distributions, we can write this as

\[
E[|x - y|] = \int_0^\infty \int_0^y (y - x) f(x) f(y) \, dx \, dy.
\]

Let this inner integral be \( I \). Then

\[
I(y) = y C(y) - \int_0^y x f(x) \, dx,
\]

\[
= y C(y) - \left\{ \int_0^y \frac{d}{dx} [x C(x)] \, dx - \int_0^y C(x) \, dx \right\},
\]

\[
\leq y C(y).
\]

Therefore, we have
\[ E[|x - y|] \leq 2 \int_0^\infty f(y) \ y \ C(y) \ dy, \]
\[ = 2 \int_0^1 y \ C \ dC, \]
\[ = \frac{2}{\log(1 - p)} \int_0^1 \log \left( 1 - C^{1/q} \right) \ dC, \]
\[ = \left[ \log(1 - p) \right] = H(2q). \]  

(C3)

where \( H(n) = \sum_{m=1}^n 1/m \) is the \( n \)th harmonic number. Here, \( q \) is equal to the total number of elementary pairs that need to connect in each “section” at a given level, which is given by \( 2^{N_S - t} \), which arrives at Eq. (33).
[39] S. Nauerth, M. Fürst, T. Schmitt-Manderbach, H. Weier, and H. Weinfurter, New Journal of Physics 11, 065001 (2009).
[40] S. E. Vinay and P. Kok, Physical Review A 97, 042335 (2018).
[41] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, Physical Review Letters 76, 722 (1996).
[42] B. C. Arnold, R. A. Groeneveld, et al., The Annals of Statistics 7, 220 (1979).
[43] H. Krauter, C. A. Muschik, K. Jensen, W. Wasilewski, J. M. Petersen, J. I. Cirac, and E. S. Polzik, Physical Review Letters 107, 080503 (2011).
[44] T. D. Ladd, D. Maryenko, Y. Yamamoto, E. Abe, and K. M. Itoh, Physical Review B - Condensed Matter and Materials Physics 71, 1 (2005).
[45] K. Nemoto, M. Trupke, S. J. Devitt, B. Scharfenberger, K. Buczak, J. Schmiedmayer, and W. J. Munro, Scientific Reports 6, 26284 (2016).
[46] J. G. Kemeny and J. L. Snell, Finite Markov chains: with a new appendix "Generalization of a fundamental matrix" (Springer, 1983).