npj | quantum information

Article

Published in partnership with The University of New South Wales

https://doi.org/10.1038/s41534-025-01000-5

Networking quantum networks with minimum cost aggregation

Koji Azuma[1,2]

A quantum version of an internet, called a quantum internet, holds promise for achieving distributed quantum sensing and large-scale quantum computer networks, as well as quantum communication among arbitrary clients all over the globe. The main building block is efficient distribution of entanglement—entangled bits (ebits)—between arbitrary clients in a quantum network with error bounded by a constant, irrespective of their distance. In practice, this should be accomplished across multiple self-organising quantum networks, analogously to what the current internet does in conventional communication. Here we present a practical recipe on how to efficiently give arbitrary clients ebits with error bounded by a constant, regardless of their distance, across multiple quantum networks. This recipe is composed of two new concepts, minimum cost aggregation and network concatenation. Our recipe forms the necessary basis of designing a quantum internet protocol for networking self-organising quantum networks to make a global-scale quantum internet.

A quantum version of an internet, called a quantum internet, is the holy grail of quantum information processing, enabling the deployment of a wide range of quantum technologies and protocols—such as quantum computation, communication, and metrology—even on a global scale[1–4]. Towards realisation of such a quantum internet, there have been not only experimental developments[5–10] but also simulations[11,12] of near-term quantum networks. A question here essential to have a global-scale quantum internet is how to connect such small self-organizing quantum networks, so as to have a larger network, analogously to what the Internet Protocol does for the current internet. The goal of this paper is to present a practical solution to this problem.

In the current internet, if a client, Alice, wants to communicate with another client, Bob, an internet protocol determines the path that the data follow to travel across multiple networks from Alice to Bob. In contrast, in the quantum world, it is risky to transfer precious quantum data, such as qubits obtained by running a large-scale quantum computer long time, relying on such routing from Alice to Bob across multiple quantum networks. This is because in contrast to classical data, quantum data cannot be duplicated or amplified[13,14] and it is thus irreplaceable. Nevertheless, it is easily deteriorated during the transmission, owing to noise of channels in the networks. Against this, quantum error correction to it enables us to transmit quantum data directly even over noisy channels, like classical data in the current internet[15]. However, the allowed error thresholds to have the blessing of quantum error-correcting codes are not so high[16]. Hence, we cannot eliminate a risk of a failure, e.g., because of an accidental error in the networks beyond the threshold, leading to dissipation of irreplaceable precious quantum data in the transmission. It would thus be better for a quantum internet to work without relying on routing of precious quantum data, implying that we cannot use classical routing protocols in the current internet as they are to control a quantum internet.

In fact, a goal of a quantum internet protocol is regarded as distribution of Bell pairs, called ebits, to Alice and Bob, across multiple quantum networks, rather than direct transmission of quantum data via routing[1–4]. This is because ebits are universal resource for clients in applications of a quantum internet[2,4], including faithful transfer of quantum data from Alice to Bob by using quantum teleportation protocol[17]. In other words, ebits are merely resource and thus replaceable in contrast to precious quantum data. Besides, their existence is testable by clients themselves[18]. Hence, efficient distribution of ebits between Alice and Bob over quantum networks is deemed a main task of a quantum internet.

In the case of a linear network where Alice and Bob can be connected by a chain of quantum repeaters, we can invoke a quantum repeater protocol[16] to give them ebits with error bounded by a constant, efficiently. However, this protocol works only for linear networks and it does not tell us how we should choose such a linear network from *given* quantum networks with arbitrary topology. Recent proposals for aggregation of quantum repeaters have tried to answer this question[3,19–21]. However, they are unsatisfactory in practice, because they require point-to-point entanglement generation—the first step of the protocol—not only

a) to be run more than necessary, but also
b) to suppress the error, depending on the whole size of the networks.

[1]NTT Basic Research Laboratories, NTT Corporation, 3-1 Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198, Japan. [2]NTT Research Center for Theoretical Quantum Information, NTT Corporation, 3-1 Morinosato-Wakamiya, Atsugi, 243-0198, Japan. e-mail: koji.azuma@ntt.com

The former requirement a) indicates that they are costly to work and not energy-efficient at all, which gets worse and worse as the size of the network increases. The latter requirement b) implies that the errors in the point-to-point entanglement generation should be lowered to zero asymptotically for the size of the networks in order to present clients with ebits with error bounded by a constant, indicating unscalability of the protocols in practical scenarios. Hence, there was no practical and scalable idea to control a quantum internet.

In this paper, we present a practical recipe on how to aggregate quantum networks with arbitrary topology in order to give clients ebits with error bounded by a constant efficiently, regardless of their distance. First, we introduce the concept of minimum cost aggregation which works over a quantum network. The cost here is a general notion like a price to pay for presenting ebits between two nodes in the network, and it is minimised in the aggregation. This way, our aggregation eliminates the requirement a) above. Besides, the minimum cost aggregation is concatenated to enable arbitrary clients to have ebits with error bounded by a constant across multiple quantum networks, regardless of the overall size. This 'network concatenation' eliminates the requirement b) above, by networking quantum networks. The concatenation is a concept broader than conceptions of the nested entanglement purification protocol in the original proposal for quantum repeaters[16] and of concatenation of quantum error-correcting codes[15]. Our ideas thus form the necessary basis to control a large-scale quantum network composed of self-organising quantum networks.

Results

Framework of aggregation protocols

We begin by introducing a general framework of aggregation protocols (Fig. 1)[3]. The goal of these protocols is to present ebits between two clients (or arbitrary two nodes) in a given quantum network, by utilising the other nodes as repeater nodes. More precisely, we associate a given quantum network with a graph $G = (V, E)$ with a set V of vertices and a set E of undirected edges, where vertices $x \in V$ correspond to quantum information processing nodes in the quantum network and each undirected edge $e = \{xy\} = \{yx\} \in E$ with $x, y \in V$ specifies quantum channels between different nodes x and y as a single quantum channel $\mathcal{N}_e$. Here xy with $x, y \in V$ is used to specify a directed edge from vertex x to vertex y, while the enclosed version in curly brackets, i.e., $\{xy\}$, means the undirected edge connecting vertices x and y. We also associate two vertices $s(\in V)$ and $t(\in V)$ with the two clients and the other vertices $r \in V\backslash\{s, t\}$ with repeater nodes. Besides, we assume that there is a maximum number $m_e^{\max}$ of uses of each channel $\mathcal{N}_e$, for instance, because of the availability of the channel for a given time interval T. In particular, $m_e^{\max}$ is assumed to be constant, at least, for this time interval T, but it can be updated after the time T in general, reflecting the traffic.

An aggregation protocol starts by running a point-to-point entanglement generation scheme, perhaps equipped with quantum error correction or entanglement distillation, over quantum channel $\mathcal{N}_e$ for every $e \in E$, each of which can provide a state $\hat{\rho}_e$ close to f_e copies of a Bell pair $|\Psi^+\rangle_e := (|01\rangle_e + |10\rangle_e)/\sqrt{2}$ for computation basis states $\{|0\rangle, |1\rangle\}$, that is, $\| \hat{\rho}_e - |\Psi^+\rangle\langle\Psi^+|_e^{\otimes f_e} \|_1 \le \delta_e$ with a given $\delta_e > 0$ (in terms of the trace distance), by using the quantum channel $\mathcal{N}_e$ $m_e (\le m_e^{\max})$ times (i.e., using $\mathcal{N}_e^{\otimes m_e}$) and local operations and classical communication (LOCC). Thus, we can have a state $\hat{\rho}_E := \bigotimes_{e\in E}\hat{\rho}_e$ close to Bell-pair network $\hat{\Psi}_E^f := \bigotimes_{e\in E}|\Psi^+\rangle\langle\Psi^+|_e^{\otimes f_e}$. Notice that f_e is a monotonically non-decreasing function of m_e for a fixed point-to-point entanglement generation scheme. Thus, if we write its maximum as $c_e := f_e(m_e^{\max})$, we should have

$$0 \le f_e \le c_e \tag{1}$$

for any m_e and $e \in E$.

For a fixed entanglement generation scheme, in general, $\| \hat{\rho}_e - |\Psi^+\rangle\langle\Psi^+|_e^{\otimes f_e} \|$ can depend not only on f_e but also on m_e, for instance, because of statistical fluctuation caused by a probabilistic process in the scheme. This means that for a given constant $\delta_e > 0$, the assumption $\| \hat{\rho}_e - |\Psi^+\rangle\langle\Psi^+|_e^{\otimes f_e} \| \le \delta_e$ (as well as $m_e \le m_e^{\max}$) restricts possible choices of m_e among ones to give f_e ebits. The smallest m_e among these possible choices for every f_e composes a function of f_e, whose inverse function can be regarded as $f_e(m_e)$ (for a given δ_e) written above.

Existing protocols

In the existing aggregation protocols[3,19–21], all the point-to-point entanglement generation schemes are assumed to use their own channel $\mathcal{N}_e$ the *maximum* number $m_e^{\max}$ of times, to establish a state close to a *maximal* Bell-pair network $\hat{\Psi}_E^c = \bigotimes_{e\in E}|\Psi^+\rangle\langle\Psi^+|_e^{\otimes c_e}$. Then, it is transformed to Bell pairs between two clients s and t by performing entanglement swapping all over distinct st-paths—paths whose ends are vertices s and t—in a multigraph where each edge corresponds to a Bell pair in the Bell-pair network $\hat{\Psi}_E^c$. According to Menger's theorem[22,23], there are

$$C_{\{st\}} := \min_{V_{s;t}(\subset V)} \sum_{e\in\partial(V_{s;t})} c_e \tag{2}$$

distinct st-paths[19] in the multigraph, where $\partial(X)$ for a subset X of V is defined as the set of undirected edges connecting a node in X and a node in $V\backslash X$ and the minimisation is taken over all subsets $V_{s;t}$ of V that include node s (i.e., $s \in V_{s;t}$) but do not node t (i.e., $t \notin V_{s;t}$ or $t \in V\backslash V_{s;t}$). As a result, those protocols provide the two clients s and t with a state close to $C_{\{st\}}$ copies of a Bell pair[3,19,21]. However, as stated as the requirement a) in the introduction, these protocols are demanding as they require us to prepare a state close to the maximal Bell-pair network $\hat{\Psi}_E^c$.

Minimum cost aggregation

The goal of our protocol is to provide a state close to an arbitrary number $F_{\{st\}}^*$ of ebits, satisfying

$$0 \le F_{\{st\}}^* \le C_{\{st\}}, \tag{3}$$

to two clients s and t. Here $F_{\{st\}}^*$ represents a demand from clients s and t in the time interval T. Notice that our protocol can provide the same number of ebits as the existing protocols[3,19–21] just by choosing $F_{\{st\}}^* = C_{\{st\}}$. However, in striking contrast to the existing protocols[3,19–21], our protocol accomplishes this goal without the necessity of preparing a state close to the maximal Bell-pair network $\hat{\Psi}_E^c$. In particular, the protocol starts by preparing a state close to a *minimal* Bell-pair network $\hat{\Psi}_{E^*}^{f^*}$, rather than the maximal one $\hat{\Psi}_E^c$, where f^* is chosen so as to minimise a total cost

$$\$^f := \sum_{e\in E} \$_e f_e, \tag{4}$$

and $E^* \subset E$ is the set of all undirected edges with $f_e^* > 0$. Here $\$_e$ is a cost to produce a unit of f_e—i.e., an ebit—on each edge e, which is assumed to be represented by a positive rational constant.

Since the cost $\$_e f_e$ for each edge $e \in E$ is proportional to f_e and thus monotonically non-decreasing also for m_e, the sufficiency of the preparation of the minimal Bell-pair network $\hat{\Psi}_{E^*}^{f^*}$ implies that the required channel uses $\{m_e\}_{e\in E}$ and related costs (e.g., device uses) are all minimised. For example, if we use entanglement generation schemes whose efficiencies $g_e := f_e/m_e$ can be deemed constant, by choosing the constant overhead g_e^{-1} as a cost $\$_e$ (i.e., $\$_e = 1/g_e$), $\$^f = \sum_{e\in E} f_e/g_e = \sum_{e\in E} m_e$ represents the total number of channel uses in the protocol explicitly; indeed, there are such point-to-point entanglement generation schemes, such as heralded entanglement generation protocol[4] over a lossy bosonic channel and ones equipped with quantum error-correcting codes[24,25] or entanglement distillation[26] with constant overheads. But, notice that this is not the only choice of the cost $\$_e$; for example, one could simply regard the cost $\$_e$ as a price of a Bell pair produced by an entanglement generation scheme over channel $\mathcal{N}_e$ (although the price $\$_e$ might be set on considering not only the number m_e of channel uses but also other factors such as profit and energy consumption).

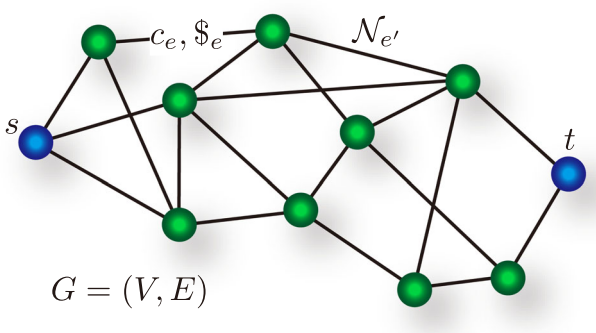

Fig. 1 | Quantum network associated with graph $G = (V, E)$. Vertices $x \in V$ correspond to quantum information processing nodes and each undirected edge $e = \{xy\} = \{yx\} \in E$ specifies quantum channels between different nodes x and y as a single quantum channel $\mathcal{N}_e$. Two vertices $s(\in V)$ and $t(\in V)$ are held by two clients, and the other vertices $x \in V\backslash\{s, t\}$ are regarded as repeater nodes. If we run a point-to-point entanglement generation scheme by using quantum channel $\mathcal{N}_e$ $m_e(\leq m_e^{\max})$ times, it is assumed to present a state $\hat{\rho}_e$ δ_e-close to $f_e(\leq c_e)$ ebits, where f_e is a monotonically non-decreasing function of m_e with $c_e := f_e(m_e^{\max})$. $\$_e$ is a cost to produce such an ebit on each edge e.

To establish our protocol, we associate the current problem with the minimum cost flow in a directed graph[27], rather than Menger's theorem[22,23] for an undirected graph, known in graph theory. For that, we first notice an equivalence[21] between a flow of qubits and share of ebits given by quantum teleportation[16]. The teleportation allows us to transmit $f_{\{xy\}}$ qubits in total by consuming $f_{\{xy\}}$ ebits, either from node $x \in V$ to node $y \in V$, i.e., to have a directed flow f_{xy} with $f_{xy} = f_{\{xy\}}$ or from node y to node x, i.e., to have a directed flow f_{yx} with $f_{yx} = f_{\{xy\}}$, implying $f_{xy} + f_{yx} = f_{\{xy\}}$ in general. Conversely, if we can send f_{xy} qubits from node x to node y faithfully, we can share $f_{\{xy\}}$ ebits between the nodes. Thanks to this equivalence, for the given graph $G(V, E)$, we can consider an induced digraph $D(V, A)$ which has the same set V of vertices as the graph G but has the set A of directed edges xy and yx induced by all undirected edges $\{xy\} \in E$.

In particular, Eq. (1) implies that the possible flows $f_{xy}(\geq 0)$ and $f_{yx}(\geq 0)$ of qubits, made by the quantum teleportation using $f_{\{xy\}}$ ebits, follow a capacity constraint,

$$0 \leq f_{xy} + f_{yx} \leq c_{\{xy\}} \tag{5}$$

for any $xy \in A$, $yx \in A$ and $\{xy\} \in E$. Besides, we require all the repeater nodes $r \in V\backslash\{s, t\}$ to make a flow f of the quantum teleportation in a way satisfying a conservation condition,

$$\sum_{a \in \partial^+(\{r\})} f_a = \sum_{a \in \partial^-(\{r\})} f_a, \tag{6}$$

where $\partial^+(X)$ ($\partial^-(X)$) for $X \subset V$ is defined as the set of directed edges in A whose tails belong to X ($V\backslash X$) and whose heads belong to $V\backslash X$ (X). The net flow F_{st}^f of qubits from node s to node t is described by

$$F_{st}^f := \sum_{a \in \partial^+(\{s\})} f_a - \sum_{a \in \partial^-(\{s\})} f_a. \tag{7}$$

If we maximise this net flow F_{st}^f over flows f on A under the constraints (5) and (6), the maximum value is equal to the minimum cut $C_{\{st\}}$ of Eq. (2), as stated in the max-flow min-cut theorem[23,27–29].

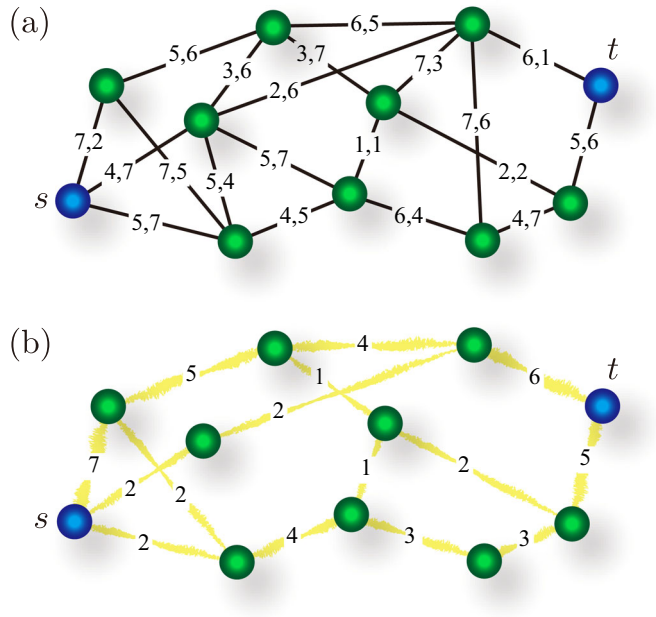

Fig. 2 | Minimum cost flow. As an example, suppose that two values on each edge e in the graph (**a**) represent c_e, $\$_e$, respectively. Then, a solution f^* of the minimum-cost maximum flow problem (that is, with assuming $F_{\{st\}}^* = C_{\{st\}}$) is described by values f_e^* on edges e in the graph (**b**). This solution shows $F_{\{st\}}^* = C_{\{st\}} = 11$ and $\$^{f^*} = 215$. This cost $\$^{f^*} = 215$ corresponds to the one which is needed to establish $F_{\{st\}}^* = C_{\{st\}} = 11$ ebits between clients s and t with our protocol. In contrast, the existing protocols[19–21] cost $\$^c = 513$ to serve the clients with the same number of ebits because they begin by generating the maximal Bell-pair network $\hat{\Psi}_E^c$.

On the other hand, since the cost for each edge $\{xy\}$ is regarded as $\$_{\{xy\}} f_{\{xy\}} = \$_{\{xy\}} f_{xy} + \$_{\{xy\}} f_{yx}$, the total cost of Eq. (4) is rephrased as

$$\$^f = \sum_{a \in A} \$_{\{a\}} f_a. \tag{8}$$

A solution of the linear program to minimise this total cost $\$^f$ over flows f on A under the net-flow constraint,

$$F_{st}^f = F_{\{st\}}^*, \tag{9}$$

for a given integer $F_{\{st\}}^*$ with Eq. (3), as well as constraints (5) and (6), is called a minimum cost flow, denoted by f^*. The solution f^* for the choice of $F_{\{st\}}^* = C_{\{st\}}$ is especially called a minimum-cost maximum flow. Depending on assumptions, there are a variety of algorithms to derive the minimum cost flows, which are computationally efficient, at least, in the case where capacities $\{c_a\}_{a \in A}$, costs $\{\$_a\}_{a \in A}$ and demand $F_{\{st\}}^*$ are integral (see ref. 27 for more general cases). Although costs $\{\$_a\}_{a \in A}$ are allowed to be rational numbers in our case, the algorithms work efficiently even in this case because we can always transform rational numbers to integers by multiplying them by a suitably large number. Figure 2 presents an example of a solution of the minimum-cost maximum-flow problem.

Now we can introduce our protocol, called minimum cost aggregation, to provide a state close to $F_{\{st\}}^*(\leq C_{\{st\}})$ ebits to the clients s and t with minimum cost. 1) The first step of this protocol is to derive a solution f^* of the minimum-cost flow problem on A for a given $F_{\{st\}}^*$. We define f^* on E as $f_{\{xy\}}^* := f_{xy}^* + f_{yx}^*$ for any $xy \in A$ and $yx \in A$. 2) Then, we run the point-to-point entanglement generation scheme over quantum channel $\mathcal{N}_e$ for every $e \in E$, to obtain a state $\hat{\rho}_e$ close to f_e^* copies of a Bell pair $|\Psi^+\rangle_e$, that is, $\| \hat{\rho}_e - |\Psi^+\rangle\langle\Psi^+|_e^{\otimes f_e^*} \|_1 \leq \delta_e$ with a given $\delta_e > 0$, by using the quantum channel $\mathcal{N}_e$ $m_e^*(\leq m_e^{\max})$ times and LOCC, where m_e^* is defined as one satisfying $f_e^* = f_e(m_e^*)$ (or the smallest m_e satisfying $f_e(m_e) > f_e^*$ if there is no solution of m_e satisfying $f_e^* = f_e(m_e)$). 3) We then

ask repeater nodes $r \in V\backslash\{s,t\}$ to perform Bell measurements [followed by local Pauli correction by client t only for simplicity of the following analysis (although this Pauli correction is unnecessary for the purpose of distribution of ebits between clients s and t like in our current case)] that would be needed to make the quantum teleportation flow f^* on A for generating the net flow $F^*_{\{st\}}$ of qubits by consuming Bell-pair network $\hat{\Psi}^{f^*}_{E^*}$. This LOCC operation works as entanglement swapping if it is applied to Bell-pair network $\hat{\Psi}^{f^*}_{E^*}$. The overall description of the protocol is summarised in the Methods.

If the LOCC operation in step 3) is a noiseless operation denoted by $\Lambda^{f^*}_{E^*}$ (satisfying $\Lambda^{f^*}_{E^*}(\hat{\Psi}^{f^*}_{E^*}) = |\Psi^+\rangle\langle\Psi^+|^{\otimes F^*_{\{st\}}}_{\{st\}}$), the state $\hat{\omega}_{\{st\}}$ served to clients s and t after the application of the LOCC operation $\Lambda^{f^*}_{E^*}$ to the true initial state $\hat{\rho}_{E^*} := \bigotimes_{e\in E^*}\hat{\rho}_e$ satisfies

$$\begin{aligned}&\left\|\hat{\omega}_{\{st\}} - |\Psi^+\rangle\langle\Psi^+|^{\otimes F^*_{\{st\}}}_{\{st\}}\right\|_1 = \left\|\Lambda^{f^*}_{E^*}(\hat{\rho}_{E^*}) - \Lambda^{f^*}_{E^*}(\hat{\Psi}^{f^*}_{E^*})\right\|_1\\ &\le \left\|\hat{\rho}_{E^*} - \hat{\Psi}^{f^*}_{E^*}\right\|_1 \le \sum_{e\in E^*}\delta_e =: \delta_{E^*}.\end{aligned} \tag{10}$$

Even if the LOCC operation in step 3) is merely a noisy operation $\Gamma^{f^*}_{E^*}$ rather than noiseless one $\Lambda^{f^*}_{E^*}$, the state $\hat{\omega}_{\{st\}}$ presented to clients s and t after the LOCC operation $\Gamma^{f^*}_{E^*}$ satisfies

$$\begin{aligned}&\left\|\hat{\omega}_{\{st\}} - |\Psi^+\rangle\langle\Psi^+|^{\otimes F^*_{\{st\}}}_{\{st\}}\right\|_1 = \left\|\Gamma^{f^*}_{E^*}(\hat{\rho}_{E^*}) - \Lambda^{f^*}_{E^*}(\hat{\Psi}^{f^*}_{E^*})\right\|_1\\ &\le \left\|\Gamma^{f^*}_{E^*}(\hat{\rho}_{E^*}) - \Gamma^{f^*}_{E^*}(\hat{\Psi}^{f^*}_{E^*})\right\|_1 + \left\|\Gamma^{f^*}_{E^*}(\hat{\Psi}^{f^*}_{E^*}) - \Lambda^{f^*}_{E^*}(\hat{\Psi}^{f^*}_{E^*})\right\|_1\\ &\le \delta_{E^*} + \epsilon_{E^*},\end{aligned} \tag{11}$$

where $\epsilon_{E^*} := \left\|\Gamma^{f^*}_{E^*}(\hat{\Psi}^{f^*}_{E^*}) - \Lambda^{f^*}_{E^*}(\hat{\Psi}^{f^*}_{E^*})\right\|_1$.

On the other hand, since f^* on A is a minimum cost flow, our protocol merely needs the minimum cost $\$^{f^*} = \sum_{a\in A}\$_{\{a\}}f^*_a$. Therefore, with consuming this minimum cost $\$^{f^*}$, our protocol gives two clients s and t a state $\hat{\omega}_{\{st\}}$ $(\delta_{E^*}+\epsilon_{E^*})$-close to $F^*_{\{st\}}$ ebits.

The demand $F^*_{\{st\}}$ in the time interval T can be chosen arbitrarily as long as it satisfies Eq. (3). Thus, clients s and t who do not rush to have ebits might choose $F^*_{\{st\}}$ so as to minimise a unit price of the ebits, $\$^{f^*}/F^*_{\{st\}}$. Notice that our minimum cost aggregation with the choice of $F^*_{\{st\}} = C_{\{st\}}$ provides the same number of Bell pairs to the clients s and t as the existing protocols[19–21] although it works with minimum cost in contrast. See an example in Fig. 2 to infer how much economical our protocol is, compared with the existing protocols[19–21]. Besides, similar to the existing protocol[19], in an asymptotic limit, our protocol achieves not only quantum but also private capacities (per time T) of any quantum network composed of distilable channels[30] (whose relative entropies equal to the quantum capacities) [despite the fact that the private capacity can be larger than the quantum capacity in general, reflecting that sharing private bits (called pbits) is a weaker task than sharing ebits[31]] (see Methods).

Network concatenation

As shown above, the minimum cost aggregation in a given network $G = (V,E)$ gives clients $s(\in V)$ and $t(\in V)$ a state $\hat{\omega}_{\{st\}}$ $(\delta_{E^*}+\epsilon_{E^*})$-close to $F^*_{\{st\}}$ ebits in a time interval T, by using the minimum cost $\$^{f^*}$. However, the errors δ_{E^*} and ϵ_{E^*} depend on the size of the network through the dependence on $|E^*|$, similar to the existing protocol[19] (although they are not larger than the errors in the existing protocol[19]). This implies that the point-to-point entanglement generation schemes, as well as LOCC operations for the entanglement swapping in step 3), should make the errors zero asymptotically for the size of the network in order to provide ebits with error bounded by a constant to clients, as stated as the requirement b) in the introduction. To overcome this point, we invoke the concept of concatenation[15] lying in the original proposal[16] of quantum repeaters. The main idea is to regard our aggregation to serve entanglement $\hat{\omega}_{\{st\}}$ to two clients s and t as an entanglement generation scheme on an undirected edge $\{st\} \in \mathcal{E}$ in a one level higher graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$, to which further minimum cost aggregation can be applied.

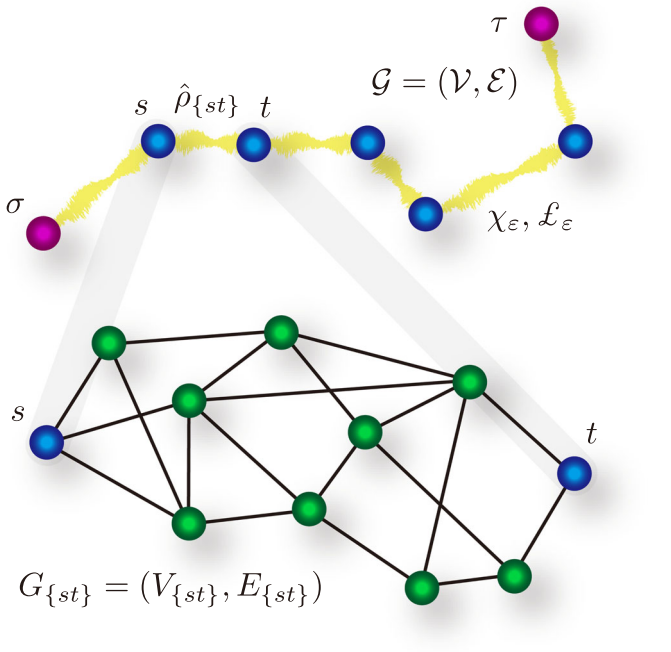

Fig. 3 | Network concatenation. For every edge $\varepsilon \in \mathcal{E}$ in a one level higher graph network $\mathcal{G} = (\mathcal{V},\mathcal{E})$, an entangled state $\hat{\rho}_\varepsilon$ δ_ε-close to $\phi_\varepsilon(\le\chi_\varepsilon)$ ebits can be obtained if we run a minimum cost aggregation protocol over its subnetwork $G_\varepsilon = (V_\varepsilon, E_\varepsilon)$ $\mu_\varepsilon(\le\mu^{\max}_\varepsilon)$ times, where ϕ_ε is a monotonically non-decreasing function of μ_ε with $\chi_\varepsilon := \phi_\varepsilon(\mu^{\max}_\varepsilon)$. $\pounds_\varepsilon$ is a cost to produce such an ebit on each edge ε. Based on these entanglement generations over edges $\varepsilon \in \mathcal{E}$, we run the minimum cost aggregation over the graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ to present ebits between nodes $\sigma \in \mathcal{V}$ and $\tau \in \mathcal{V}$. Further, these ebits themselves may be regarded as entanglement between nodes belonging to an even higher level graph network.

To clarify such a 'network concatenation' (see Fig. 3), suppose that every undirected edge $\varepsilon \in \mathcal{E}$ of a graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ specifies a one level lower graph $G_\varepsilon = (V_\varepsilon, E_\varepsilon)$ corresponding to a quantum network in which our minimum cost aggregation protocol can give the two-end vertices of ε a state $\hat{\omega}_\varepsilon$ $(\delta_{E^*_\varepsilon}+\epsilon_{E^*_\varepsilon})$-close to F^*_ε copies of a Bell pair $|\Psi^+\rangle_\varepsilon$ in a time interval T_ε by putting the minimum cost $\$^{f^*}_\varepsilon$, where $\mathcal{V}$ is the set composed of all the two ends of edges $\varepsilon \in \mathcal{E}$. In other words, the graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ describes a network of networks associated with graphs $\{G_\varepsilon\}_{\varepsilon\in\mathcal{E}}$, in which the two-end vertices of an edge $\varepsilon \in \mathcal{E}$ can be connected by entanglement $\hat{\omega}_\varepsilon$ if we run the minimum cost aggregation on the network G_ε in a time interval T_ε. Here we assume that quantum networks G_ε can work independently with each other and there is a maximum number $\mu^{\max}_\varepsilon$ of uses of each quantum network G_ε. Besides, we assume that $\delta_{E^*_\varepsilon} + \epsilon_{E^*_\varepsilon}$ is small enough (at least, less than the threshold) to run an entanglement distillation protocol over every edge $\varepsilon \in \mathcal{E}$. This way, if we make $\mu_\varepsilon(\le\mu^{\max}_\varepsilon)$ copies of the state $\hat{\omega}_\varepsilon$ by running the minimum cost aggregation μ_ε times (i.e., by taking time $\mu_\varepsilon T_\varepsilon$), the distillation protocol can convert the state $\hat{\omega}^{\otimes\mu_\varepsilon}_\varepsilon$ to a state $\hat{\rho}_\varepsilon$ close to ϕ_ε copies of a Bell pair $|\Psi^+\rangle_\varepsilon$ with a given error $\delta_\varepsilon > 0$, that is, $\|\hat{\rho}_\varepsilon - |\Psi^+\rangle\langle\Psi^+|^{\otimes\phi_\varepsilon}_\varepsilon\|_1 \le \delta_\varepsilon$. Here note that if we use the pumping protocol[16] for the entanglement distillation, the assumption of making copies of $\hat{\omega}_\varepsilon$ is not necessary, adapting the protocol to cases where network parameters for G_ε such as capacities, errors and costs may be updated every time interval T_ε for the distillation

time. Those imply that we can obtain a state $\hat{\rho}_{\mathcal{E}} := \bigotimes_{\varepsilon\in\mathcal{E}}\hat{\rho}_{\varepsilon}$ close to Bell-pair network $\hat{\Psi}_{\mathcal{E}}^{\phi} := \bigotimes_{\varepsilon\in\mathcal{E}}|\Psi^{+}\rangle\langle\Psi^{+}|_{\varepsilon}^{\otimes\phi_{\varepsilon}}$ by using every quantum network G_{ε} $\mu_{\varepsilon}(\le\mu_{\varepsilon}^{\max})$ times (to have $\hat{\omega}_{\varepsilon}^{\otimes\mu_{\varepsilon}}$) and by combining it with an entanglement distillation protocol. Similar to f_e in the above minimum cost aggregation, ϕ_{ε} is considered to be a monotonically non-decreasing function of μ_{ε} satisfying

$$0 \le \phi_{\varepsilon} \le \chi_{\varepsilon} \tag{12}$$

with $\chi_{\varepsilon} := \phi_{\varepsilon}(\mu_{\varepsilon}^{\max})$, for any μ_{ε} and $\varepsilon \in \mathcal{E}$.

Besides, suppose that a positive rational cost to produce a unit of ϕ_{ε} over the graph G_{ε} is written as $\pounds_{\varepsilon}$. As a result, the total cost of the aggregation over graph $\mathcal{G}$ is

$$\pounds^{\phi} := \sum_{\varepsilon\in\mathcal{E}} \pounds_{\varepsilon}\phi_{\varepsilon}. \tag{13}$$

If the efficiency per cost $\gamma_{\varepsilon} := \phi_{\varepsilon}/(\mu_{\varepsilon}\$_{\varepsilon}^{f^*})$ of the generation of the distilled entanglement $\hat{\rho}_{\varepsilon}$ over graph G_{ε} can be assumed to be constant over time interval $\max_{\varepsilon\in\mathcal{E}}\mu_{\varepsilon}T_{\varepsilon}$, by choosing the constant overhead $\gamma_{\varepsilon}^{-1}$ as a cost $\pounds_{\varepsilon}$ (i.e., $\pounds_{\varepsilon} = 1/\gamma_{\varepsilon}$), the total cost $\pounds^{\phi}$ can be associated with original costs $\{\$_{\varepsilon}^{f^*}\}_{\varepsilon\in\mathcal{E}}$ as $\pounds^{\phi} = \sum_{\varepsilon\in\mathcal{E}}\phi_{\varepsilon}/\gamma_{\varepsilon} = \sum_{\varepsilon\in\mathcal{E}}\mu_{\varepsilon}\$_{\varepsilon}^{f^*}$. Again, however, this is not the only choice of the cost $\pounds_{\varepsilon}$; $\pounds_{\varepsilon}$ may be a price to buy a unit of ϕ_{ε} over the network G_{ε}, set by taking into account all the necessary operations and various factors, including even additional workloads of the lower level aggregation associated with solving the minimum-cost flow problem in the step 1) and performing entanglement swapping in the step 3).

Then, we notice that the minimum cost aggregation over a graph $G = (V, E)$ introduced above can work for two clients $\sigma \in \mathcal{V}$ and $\tau \in \mathcal{V}$ even over the one level higher graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the following substitutions:

$$\begin{aligned} G(V,E) &\mapsto \mathcal{G}(\mathcal{V},\mathcal{E}),\\ m_e(\le m_e^{\max}) &\mapsto \mu_{\varepsilon}(\le\mu_{\varepsilon}^{\max}),\\ f_e(\le c_e) &\mapsto \phi_{\varepsilon}(\le\chi_{\varepsilon}),\\ C_{\{st\}} &\mapsto X_{\{\sigma\tau\}},\\ \$_e &\mapsto \pounds_{\varepsilon}, \end{aligned} \tag{14}$$

where

$$X_{\{\sigma\tau\}} := \min_{\mathcal{V}_{\sigma;\tau}(\subset\mathcal{V})} \sum_{\varepsilon\in\partial(\mathcal{V}_{\sigma;\tau})} \chi_{\varepsilon} \tag{15}$$

minimised over subsets $\mathcal{V}_{\sigma;\tau}$ of $\mathcal{V}$ that include node σ (i.e., $\sigma \in \mathcal{V}_{\sigma;\tau}$) but do not node τ (i.e., $\tau \notin \mathcal{V}_{\sigma;\tau}$ or $\tau \in \mathcal{V}\setminus\mathcal{V}_{\sigma;\tau}$). In particular, for any two clients $\sigma \in \mathcal{V}$ and $\tau \in \mathcal{V}$ and for a given integer demand $\Phi_{\{\sigma\tau\}}^{*}$ with

$$0 \le \Phi_{\{\sigma\tau\}}^{*} \le X_{\{\sigma\tau\}}, \tag{16}$$

based on a derived minimum cost flow ϕ^{*}, the minimum aggregation protocol over the one level higher network $\mathcal{G}$ gives two clients σ and τ state $\hat{\omega}_{\{\sigma\tau\}}$ $(\delta_{\mathcal{E}^*} + \epsilon_{\mathcal{E}^*})$-close to $\Phi_{\{\sigma\tau\}}^{*}$ copies of a Bell pair $|\Psi^{+}\rangle_{\{\sigma\tau\}}$ by using the minimum cost $\pounds^{\phi^*}$, where $\mathcal{E}^*$ represents the set of edges $\varepsilon \in \mathcal{E}$ with $\phi_{\varepsilon}^{*} > 0$ and

$$\delta_{\mathcal{E}^*} := \sum_{\varepsilon\in\mathcal{E}^*} \delta_{\varepsilon}, \tag{17}$$

$$\epsilon_{\mathcal{E}^*} := \left\| \Gamma_{\mathcal{E}^*}^{\phi^*}(\hat{\Psi}_{\mathcal{E}^*}^{\phi^*}) - \Lambda_{\mathcal{E}^*}^{\phi^*}(\hat{\Psi}_{\mathcal{E}^*}^{\phi^*}) \right\|_1 \tag{18}$$

are defined analogously to Eqs. (10) and (11). Since δ_{ε} could be chosen arbitrarily (as long as there is an entanglement distillation protocol as required), the error $\delta_{\mathcal{E}^*} + \epsilon_{\mathcal{E}^*}$ depends only on the size of the current graph $\mathcal{G}$, independent of those of one level lower graphs G_{ε}. This is an advantage of concatenating aggregation.

Repeatedly, if the error $\delta_{\mathcal{E}^*} + \epsilon_{\mathcal{E}^*}$ of the current aggregation is small enough (at least, less than a threshold) to perform an entanglement distillation protocol (which can transform copies of the state $\hat{\omega}_{\{\sigma\tau\}}$ to a state close to fewer copies of a Bell pair $|\Psi^{+}\rangle_{\{\sigma\tau\}}$), we can consider further aggregation over an even higher level of a graph network by regarding the current aggregation as an entanglement generation scheme on an undirected edge $\{\sigma\tau\}$ in the higher-level graph. Therefore, by properly concatenating the aggregation, we can provide ebits with an error bounded by a constant to any pair of clients across multiple quantum networks with arbitrary topology, independently of the whole size.

Notice that the idea of the network concatenation is reduced to the conception of the nested entanglement purification protocol introduced by the original proposal for quantum repeaters[16], stemming from the idea of the concatenation of quantum error-correcting codes[15]. Indeed, this reduction is possible just by applying our network concatenation idea to linear networks, sequentially.

Discussion

One might wish to have a global optimal protocol which minimises a total cost upon considering all the information about the whole network. This is of fundamental importance but may not be practical if we see how the current internet grows, in which individual internet providers appear and build own networks and thus no one knows even the structure of the whole network. Therefore, in the future, it would be natural that even at the expense of global optimality, the minimum cost aggregation is applied to quantum subnetworks, independently, as subnetworks may be organized and governed by different network holders, such as individuals, public, institutes, universities, companies, and governments, like the current internet. Then, the idea of the network concatenation would play the role analogous to the current Internet, that is, to connect subnetworks to have 'a network of networks', although they differ in what they serve: the former distributes resource in the form of correlation, while the latter transmits data itself.

We have only just begun to grasp a possible form of the future quantum internet: before concluding this paper, let us discuss possible generalisations of our introduced concepts. In this paper, our concepts have been introduced by considering quantum networks spread over space with static specification of network parameters such as capacities, errors and costs for a given time interval. However, in practice, the network parameters to run the minimum cost aggregation in a level of a network should be updated every the time interval, reflecting the change in the traffic owing to, for example, appearance of more pairs of clients and occurrence of burst errors in the quantum networks. If this update can be made frequently enough compared with the change in the traffic, our protocol can still work, including even the network concatenation with the use of the pumping protocol[16] for the entanglement distillation. Even if not and our protocol then does not work properly, however, notice that it does not cause a damage which would entail compensation, because it distributes ebits between clients, that is, merely self-testable resource for quantum communication.

Besides, in our paper, the cost for each edge in the minimum cost aggregation is assumed to be linear to the number of ebits (that is, the flow) served by the edge, which renders the underlying minimum-cost flow problem a linear program. However, one might wish to regard the cost for each edge as a nonlinear function of the number of ebits supplied by the edge. Even in such a general case, in graph theory, there are several classes of minimum-cost flow problems to be solvable efficiently[27], including practical scenarios where the cost for each edge is convex for the flow, implying that our formulation can also be generalised accordingly. For instance, even if one might wish to regard the number m_e of channel uses (or network uses) as the cost, the underlying minimum-cost flow problem could be solved efficiently in the case where $f(m_e)$ of the entanglement generation scheme for each edge e is a concave function of m_e. Or, perhaps, upon considering that

error δ_e of generated ebits can depend on the number f_e of ebits in general, one might generalise our formulation so as to be able to choose the error δ_e as the cost, to minimise the total error δ_{E^*} as a quality control of served ebits to clients. By solving this problem, a network provider could identify a subnetwork which can provide ebits with a bounded error to clients, in advance.

Our protocol can be generalised to work even over a quantum network spread over spacetime[32] in order to include effects of time lags caused by necessary classical communication between distant sites, as well as the accompanied noise induced by quantum memories (as explored in the context of conventional quantum repeater protocols[33–38]). In this case, one might want to consider a quantum network that includes a quantum channel associated with a directed edge to reflect its directionality: for instance, a noisy quantum memory acts as a one-way quantum channel forward in time. In this kind of quantum network over spacetime consisting of quantum channels, each represented by either an undirected or a directed edge, one might consider a task[32] to distribute ebits between clients located in a time slice. However, notice that our minimum-cost aggregation protocol could work even for this case, because it is designed over a directed graph in any case.

Although our recipe is mainly discussed by focusing on the distribution of ebits to a pair of clients, it can be applied to more practical cases where multiple pairs of clients appear at the same time in a quantum network. Indeed, this can be done, at least, by applying our recipe in the form of a greedy algorithm with proper updates of network parameters. However, it is interesting to consider a question whether a combination between our idea and a quantum networking coding[39–41] leads to a more efficient solution in this scenario.

Following possible generalisation of our recipe like ones discussed here, an efficient algorithm to control a real quantum internet would appear. Since the cost introduced in this paper can be associated with the price of a Bell pair, our theory would also be the basis of pricing quantum communication services served by a quantum internet, that is, the economics of a quantum internet, which is related to one of most important open questions raised in ref. 4.

Methods

Overall description of the minimum-cost aggregation protocol

The whole description of the minimum-cost aggregation protocol can be summarised as follows. Suppose that a quantum network specified by an undirected graph $G = (V, E)$ has entanglement generation protocols over edges $e \in E$ characterised by network parameters $\{c_e, \delta_e, \$_e\}_{e\in E}$ for a time interval T, each of which can generate a state $\delta_e(\geq 0)$-close to $f_e(\leq c_e)$ ebits between the two ends of the edge e with cost $f_e\$_e(\geq 0)$. For the given network G, we introduce a directed graph $D(V, A)$ which has the same set V of vertices as the graph G but has the set of directed edges xy from x to y and yx from y to x for all undirected edges $\{xy\} \in E$. Then, the minimum-cost aggregation protocol is formally defined as follows:

0) Given a pair of targeted clients $s \in V$ and $t \in V$, we calculate the minimum cut $C_{\{st\}}$ of capacities defined by Eq. (2), which is then made public as the possible maximum amount of demand $F^*_{\{st\}}$.
1) For a given demand $F^*_{\{st\}}$ satisfying Eq. (3), we find a solution f^* which is a flow $\{f_a\}_{a\in A}$ over the directed graph $D(V, A)$ that minimises the total cost $\$^f$ defined in Eq. (8), under constraints of Eqs. (5-7) and Eq. (9). From this solution f^* on A, we define f^* on E as $f^*_{\{xy\}} := f^*_{xy} + f^*_{yx}$ for any $xy \in A$ and $yx \in A$ and E^* as the subset of undirected edges in E with $f^*_e > 0$.
2) Given the solution f^* on E, we ask the two ends of each edge $e \in E^*$ to run the entanglement generation protocol over the edge e to make a state δ_e-close to f^*_e ebits. This presents a state δ_{E^*}-close to a Bell-pair network $\hat{\Psi}^{f^*}_{E^*}$ which has $F^*_{\{st\}}$ st-paths composed of Bell pairs, where δ_{E^*} is defined in Eq. (10).
3) We finally ask repeater nodes $r \in V\backslash\{s, t\}$ to perform Bell measurements for entanglement swapping along each of the $F^*_{\{st\}}$ st-paths. This gives clients s and t a state $\hat{\omega}_{\{st\}}$ $(\delta_{E^*} + \epsilon_{E^*})$-close to $F^*_{\{st\}}$ ebits, where ϵ_{E^*} is the imperfection of the swapping defined in Eq. (11).

Steps 0) and 1) constitute the classical phase of the protocol, while steps 2) and 3) compose its quantum phase.

Notice that how much error $\delta_{E^*} + \epsilon_{E^*}$ is held by the final state $\hat{\omega}_{\{st\}}$ is predictable in the step 1). This means that if the predicted error is larger than a threshold to make a subsequent protocol (such as entanglement distillation) work, one might run step 1) again as a compromise by choosing a smaller demand $F^*_{\{st\}}$ or might consider to use a subnetwork of the given network G, rather than the whole, before running the quantum phase of steps 2) and 3). Or, more conservatively, by comparing the worst error $\delta_E + \epsilon_E$ (rather than $\delta_{E^*} + \epsilon_{E^*}$) with the threshold before running the protocol, one might choose a subnetwork of the given network G as a network to which the minimum cost aggregation is applied.

Asymptotic limit

Let us derive the ultimate performance of our protocol by taking an asymptotic limit. For that, we assume that all the local operations are noiseless and introduce a parameter T like 'time', which is arbitrary as long as rates $r_e := m^{\max}_e/T$ are constant. Then, for the fixed rate $r_e > 0$, $m^{\max}_e \to \infty$ in the limit of $T \to \infty$, for which, by optimizing the point-to-point entanglement generation schemes used in our protocol, $f_e(m^{\max}_e)/m^{\max}_e = c_e/m^{\max}_e$ could reach the quantum capacity $Q^{\leftrightarrow}(\mathcal{N}_e)$ of channel $\mathcal{N}_e$ assisted by unlimited forward and backward classical communication and the error δ_e follows $\delta_e \to 0$. Then, the number $F^*_{\{st\}}$ of ebits served to clients s and t per 'time' T in the asymptotic limit of $T \to \infty$, i.e., the asymptotic rate $R := \lim_{T\to\infty} F^*_{\{st\}}/T$ is upper bounded as

$$\begin{aligned} R &\leq \lim_{T\to\infty} \frac{C_{\{st\}}}{T} = \lim_{T\to\infty} \min_{V_{s:t}} \sum_{e\in\partial(V_{s:t})} \frac{m^{\max}_e}{T}\frac{c_e}{m^{\max}_e} \\ &\leq \min_{V_{s:t}} \sum_{e\in\partial(V_{s:t})} r_e Q^{\leftrightarrow}(\mathcal{N}_e). \end{aligned} \tag{19}$$

Clearly, this upper bound is achievable by our protocol with $F^*_{\{st\}} = C_{\{st\}}$ using the optimal point-to-point entanglement generation schemes, in which the served ebits are asymptotically noiseless according to Eq. (10). Therefore, our protocol provides $\min_{V_{s:t}} \sum_{e\in\partial(V_{s:t})} r_e Q^{\leftrightarrow}(\mathcal{N}_e)$ ebits to the clients s and t per 'time' T asymptotically, which is the same performance as the existing protocol[19] although the consumed cost in our protocol is minimum in striking contrast. As a result, for instance, similar to the existing protocol[19], our current protocol achieves the quantum and private capacities (per 'time' T) of any quantum network composed of distilable channels[30] [whose relative entropies equal to the quantum capacities] (see refs. 3,4,19,21 for other implications).

Data availability

The author declares that all the data supporting the findings of this study are available within the paper.

Received: 29 May 2024; Accepted: 23 February 2025;
Published online: 21 March 2025

References

1. Kimble, H. J. The quantum internet. *Nature* **453**, 1023–1030 (2008).
2. Wehner, S., Elkouss, D. & Hanson, R. Quantum internet: A vision for the road ahead. *Science* **362**, eaam9288 (2018).
3. Azuma, K., Bäuml, S., Coopmans, T., Elkouss, D. & Li, B. Tools for quantum network design. *AVS Quantum Sci.* **3**, 014101 (2021).
4. Azuma, K. et al. Quantum repeaters: From quantum networks to the quantum internet. *Rev. Mod. Phys.* **95**, 045006 (2023).
5. Chen, Y.-A. et al. An integrated space-to-ground quantum communication network over 4,600 kilometres. *Nature* **589**, 214–219 (2021).
6. Joshi, S. K. et al. A trusted node-free eight-user metropolitan quantum communication network. *Sci. Adv.* **6**, eaba0959 (2020).
7. Hermans, S. L. N. et al. Qubit teleportation between non-neighbouring nodes in a quantum network. *Nature* **605**, 663 (2022).

8. Stolk, A. J. et al. Metropolitan-scale heralded entanglement of solid-state qubits. *Sci. Adv.* **10**, eadp6442 (2024).
9. Knaut, C. M. et al. Entanglement of nanophotonic quantum memory nodes in a telecom network. *Nature* **629**, 573–578 (2024).
10. Liu, J.-L. et al. Creation of memory-memory entanglement in a metropolitan quantum network. *Nature* **629**, 579–585 (2024).
11. Yehia, R., Neves, S., Diamanti, E. & Kerenidis, I. Quantum City: simulation of a practical near-term metropolitan quantum network. Preprint at https://arxiv.org/abs/2211.01190.
12. Yehia, R. et al. Connecting quantum cities: simulation of a satellite-based quantum network. *New J. Phys.* **26** 073015 (2024).
13. Wootters, W. K. & Zurek, W. H. A single quantum cannot be cloned. *Nature* **299**, 802–803 (1982).
14. Dieks, D. Communication by EPR devices. *Phys. Lett. A* **92**, 271–272 (1982).
15. Knill, E. & Laflamme, R. Concatenated quantum codes. Preprint at https://arxiv.org/abs/quant-ph/9608012.
16. Briegel, H. J., Cirac, J. I., Dür, W. & Zoller, P. Quantum repeaters: The role of imperfect local operations in quantum communication. *Phys. Rev. Lett.* **81**, 5932–5935 (1998).
17. Bennett, C. H. et al. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **70**, 1895–1898 (1993).
18. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V. & Wehner, S. Bell nonlocality. *Rev. Mod. Phys.* **86**, 419–478 (2014).
19. Azuma, K. & Kato, G. Aggregating quantum repeaters for the quantum internet. *Phys. Rev. A* **96**, 032332 (2017).
20. Pirandola, S. End-to-end capacities of a quantum communication network. *Commun. Phys.* **2**, 51 (2019).
21. Bäuml, S., Azuma, K., Kato, G. & Elkouss, D. Linear programs for entanglement and key distribution in the quantum internet. *Commun. Phys.* **3**, 55 (2020).
22. Menger, K. Zur allgemeinen Kurventheorie. *Fund. Math.* **10**, 96–115 (1927).
23. Bondy, J. A. & Murty, U. S. R. *Graph Theory* (Graduate Texts in Mathematics) (Springer, London, 2008), Vol. 244.
24. Gottesman, D. Fault-tolerant quantum computation with constant overhead. *Quantum Inf. Comput.* **14**, 1338 (2014).
25. Yamasaki, H. & Koashi, M. Time-Efficient Constant-Space-Overhead Fault-Tolerant Quantum Computation. *Nat. Phys.* **20**, 247 (2024).
26. Pattison, C. A., Baranes, G., Pablo Bonilla Ataides, J., Lukin, M. D. & Zhou, H. Fast quantum interconnects via constant-rate entanglement distillation, Preprint at https://arxiv.org/abs/2408.15936.
27. Ahuja, R. K., Magnanti, T. L. & Orlin, J. B. *Network Flows: Theory, Algorithms, and Applications* (Prentice-Hall, 1993).
28. Ford Jr, L. R. & Fulkerson, D. R. Maximal flow through a network. *Canad. J. Math.* **8**, 399–404 (1956).
29. Elias, P., Feinstein, A. & Shannon, C. E. A note on the maximum flow through a network. *IRE. Trans. Inf. Theory* **2**, 117–119 (1956).
30. Pirandola, S., Laurenza, R., Ottaviani, C. & Banchi, L. Fundamental limits of repeaterless quantum communications. *Nat. Commun.* **8**, 15043 (2017).
31. Horodecki, K., Horodecki, M., Horodecki, P. & Oppenheim, J. Secure key from bound entanglement. *Phys. Rev. Lett.* **94**, 160502 (2005).
32. Azuma, K., Mizutani, A. & Lo, H.-K. Fundamental rate-loss trade-off for the quantum internet. *Nat. Commun.* **7**, 13523 (2016).
33. Razavi, M., Piani, M. & Lütkenhaus, N. Quantum repeaters with imperfect memories: cost and scalability. *Phys. Rev. A* **80**, 032301 (2009).
34. Jiang, L. et al. Quantum repeater with encoding. *Phys. Rev. A* **79**, 032325 (2009).
35. Munro, W. J., Harrison, K. A., Stephens, A. M., Devitt, S. J. & Nemoto, K. From quantum multiplexing to high-performance quantum networking. *Nat. Photon.* **4**, 792–796 (2010).
36. Fowler, A. G. et al. Surface code quantum communication. *Phys. Rev. Lett.* **104**, 180503 (2010).
37. Munro, W. J., Stephens, A. M., Devitt, S. J., Harrison, K. A. & Nemoto, K. Quantum communication without the necessity of quantum memories. *Nat. Photon.* **6**, 777–781 (2012).
38. Azuma, K., Tamaki, K. & Lo, H.-K. All-photonic quantum repeaters. *Nat. Commun.* **6**, 6787 (2015).
39. Hayashi, M. Prior entanglement between senders enables perfect quantum network coding with modification. *Phys. Rev. A* **76**, 040301(R) (2007).
40. Leung, D., Oppenheim, J. & Winter, A. Quantum network communication–the butterfly and beyond. *IEEE Transac. Inform. Theory* **56**, 3478 (2010).
41. Kobayashi, H., Gall, F. L., Nishimura, H. & Rötteler, M. Constructing quantum network coding schemes from classical nonlinear protocols. in *Proc. 2011 IEEE International Symposium on Information Theory* 109 (St. Petersburg, Russia, 2011).

Acknowledgements

We thank Stefan Bäuml, David Elkouss, Takuya Hatomura, Toshimori Honjo, Hoi-Kwong Lo, William J. Munro, and Kiyoshi Tamaki for helpful discussion. We acknowledge the support, in part, from Moonshot R&D, JST JPMJMS2061, from JSPS KAKENHI 21H05183 JP, and from R&D of ICT Priority Technology (JPMI00316).

Author contributions

The author contributed to the conception of the work, its refinement and generalization, and its presentation and writing of the present paper.

Competing interests

The author declares no competing interests

Additional information

Correspondence and requests for materials should be addressed to Koji Azuma.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.