On Utility-optimal Entanglement Routing in Quantum Networks

Sounak Kar
QuTech, TU Delft

Arpan Mukhopadhyay
University of Warwick

***Abstract*—Quantum networks are envisioned to enable reliable distribution and manipulation of quantum information across distance, forming the foundation of a future quantum internet. The fair and efficient allocation of communication resources in such networks has been addressed through the quantum network utility maximisation (QNUM) framework, which optimises network utility under the assumption of predetermined routes for competing user demands. In this work, we relax this assumption and aim to identify optimal routes that maximise the overall network utility. Specifically, we formulate the single-path utility-based entanglement routing problem as a mixed-integer convex program (MICP). The formulation is exact when negativity is chosen as the entanglement measure for utility quantification *or* the network supports sufficiently high entanglement generation rates across demands. For other entanglement measures considered, the formulation approximates the problem with over 99.99% accuracy on evaluated real-world examples. To improve computational tractability, we then propose a randomised rounding-based heuristic and an upper bound using the relaxation of the MICP. Furthermore, based on min-congestion routing, we introduce an alternative randomised heuristic and upper bound. This heuristic is computationally faster than the previous one, while both the heuristic and the upper bound often outperform their counterparts on considered real-world networks. Our work provides the framework for extending classical flow-based and quality of service-aware routing concepts to quantum networks.**

***Index Terms*—Entanglement routing, network utility maximisation (NUM), quantum network utility maximisation (QNUM), mixed-integer convex program (MICP), randomised rounding.**

I. Introduction

Quantum networks hold the promise of facilitating tasks [1]–[3] which are fundamentally beyond the reach of classical networks. The performance of a communication network at both user and network levels is typically characterised by Quality of Service (QoS) metrics that capture the needs of underlying applications. A common framework for optimising QoS at the network level is network utility maximisation (NUM) [4], [5]. NUM quantifies user satisfaction through a utility function that maps the allocated network resources to individual utilities, which are then aggregated at the network level. The network utility is subsequently maximised over all feasible resource allocations to achieve the most *fair* and *efficient* resource distribution. In contrast to classical networks, where transmission rate typically constitutes the primary communication resource, user utility in quantum networks fundamentally depends on both entanglement generation rate and fidelity. This problem is termed quantum network utility maximisation (QNUM) [6], and approaches for its efficient solution have been investigated in [7].

Notably, QNUM addresses network utility maximisation under the assumption that routes are predetermined for each demand, represented by pairs of end nodes. In this work, we relax this assumption and seek to identify routes that maximise the highest achievable network utility. In other words, our objective is to jointly optimise network utility over both feasible routes and resource allocations. Such problems have been studied in classical networks, arising naturally in the contexts of QoS routing [8], optimal network pricing [9] and opportunistic routing in wireless networks [10]. Utility-based routing can be viewed as a generalisation of network flow models [11], which have been extensively employed for diverse objectives including throughput maximisation and congestion minimisation and remain relevant in modern networks [12]. A general framework for addressing such problems was established through multipath utility maximisation [13], which requires the candidate routes for each demand to be precomputed. However, in general, the number of feasible routes for a given demand grows exponentially with network size.

In the context of quantum networks, the problem of routing entanglement has recently received significant attention. Rather than providing a comprehensive review of this expanding literature (see [14] and references therein for an exhaustive overview), we focus on summarising key approaches and clarifying how their objectives differ from ours. Ref. [15] proposes a method to minimise the number of quantum measurements required to establish an end-to-end (e2e) link by modifying cost inputs to Dijkstra's algorithm suitably, while [16] proposes a decentralised method for shortest-path routing. A majority of the literature has focused on optimisation of entanglement generation rate. For instance, [17] presents greedy strategies for maximising generation rate on 2D grids and extending to multi-demand rate region maximisation, whereas [18] applies a multi-commodity flow formulation to maximise generation rate under multiple demands, enforcing fidelity guarantees by limiting hop counts. On the fidelity side, [19] derives low-complexity purification-enabled algorithms providing fidelity guarantees. Moreover, dynamic settings have been explored in [20], addressing peak demands, aggregate rate or service delay, and in [21], aiming to satisfy the maximum number of user demands within deadlines. However, the problem of fair and efficient resource allocation via entanglement routing remains largely unexplored, a gap this work seeks to address. We specifically study the single-path utility-based entanglement routing

problem, where each demand is routed along a single path.

In a naive approach, the single-path utility-based routing problem can be solved by addressing multiple instances of the QNUM problem, each corresponding to a feasible routing for given demands, and selecting the one with the highest network utility. However, the number of feasible routes for a given demand grows exponentially with the number of network nodes n, rendering this approach computationally prohibitive. To address this, we make the following contributions:

- We formulate the utility-based routing problem as a mixed-integer convex program (MICP) with $2kl$ binary and $6kl+k+l$ variables in total, where k and l denote the number of demands and links, respectively. Following [6], we only consider secret key fraction (SKF), distillable entanglement (DE) and negativity as entanglement measures [22], the function that quantifies the *user satisfaction* for an allocated fidelity level. While the MICP is exact for negativity, or when the network supports sufficiently high entanglement generation rates per demand, we propose a method to bound the approximation error for other cases. For considered real-world examples, the error is seen to be below 0.001%.
- While the MICP remains tractable for networks of moderate size, we propose to use its convex relaxation for (i) deriving an upper bound for the maximum achievable network utility and (ii) a randomised heuristic based on the idea of randomised rounding [23] for large networks.
- We further introduce a min-congestion-based [23] randomised heuristic which is computationally faster and an associated upper bound. On evaluated real-world examples, the second randomised heuristic consistently yields superior average performance, while the upper bound is often closer to the optimum than its counterpart.

The rest of the paper is structured as follows. Sect. II introduces the network model and entangled link generation process. The general problem formulation is presented in Sect. III, followed by the direct link-based and min-congestion formulations in Sect. IV-A and IV-B, respectively. Finally, Sect. V presents numerical evaluations on real-world networks.

II. Preliminaries

In this section, we provide a brief overview of the network structure and the state description of the quantum communication links within the network. Notation-wise, we use $[m]$ to denote the set $\{1, 2, \ldots, m\}$ for $m \in \mathbb{N}$, whereas B_j and $B_{:i}$ respectively denote the jth row and ith column of a matrix B.

A. Network and demands

We represent the network topology as a graph $G := (V, E)$, where V denotes the set of nodes and E denotes the set of edges. The end nodes in V correspond to users and the edges in E represent the direct quantum communication links between adjacent nodes. We assume that we are given a set of demands D, comprising k source-destination (SD) pairs

$$D := \{(s_i, t_i) : s_i, t_i \in V,\, i \in [k]\}\,. \tag{1}$$

A route is defined as a path in G between the corresponding SD pair. We restrict routes to simple paths for each SD pair, a choice we will later show entails no loss of generality. We assume that demand (s_i, t_i) (alternatively, demand i) has p_i many routes. We index the routes sequentially and the set of route indices serving demand i is denoted as I_i. Further, We prune the edges and vertices that are not incident on any route. We update the notations G, V, E to denote this pruned graph and assume that G has r routes and l links in total. That is,

$$r = \sum_{i \in [k]} p_i\ , \quad l = |E|\ .$$

B. Link Generation and State Description

In our network graph, each edge represents a direct quantum communication link between adjacent users/repeaters. We assume that entanglements in these elementary links are produced using the single-click protocol [24], where the generated states have the following form:

$$\rho = (1-\alpha)|\Psi^+\rangle\langle\Psi^+| + \alpha|\uparrow\uparrow\rangle\langle\uparrow\uparrow|\ . \tag{2}$$

Here α denotes the bright-state population and $|\Psi^+\rangle$ is a Bell-state orthogonal to the bright state $|\uparrow\uparrow\rangle$. The probability of success of each generation attempt is given by

$$p_{\text{elem}} = 2\kappa\eta\alpha\ ,$$

where $\kappa \in (0, 1)$ is a multiplicative factor accounting for the inefficiencies other than photon loss in the fibre. Further, η denotes the transmissivity of the link from one end to the midpoint heralding station. For a link of length L km, its transmissivity can be calculated as $\eta = 10^{-0.02L}$.

Motivated by the mathematical convenience of handling Werner states and the fact that any bipartite state can be transformed into a Werner state of same fidelity [25], [26], we assume that the elementary links generated as (2) are further converted to Werner states. Accordingly, the link state can be described as

$$\rho_w = w|\Psi^+\rangle\langle\Psi^+| + (1-w)\ \mathbb{I}_4/4\ . \tag{3}$$

For the states in (2) and (3) to have same fidelity, we must have

$$1 - \alpha = \frac{1+3w}{4}\ , \quad \text{i.e.,} \quad \alpha = \frac{3(1-w)}{4}\ .$$

If entanglement generation is attempted every T unit of time, the resulting generation rate can be expressed as

$$\frac{p_{\text{elem}}}{T} = d(1-w) \ \text{ with } \ d := \frac{3\kappa\eta}{2T}\,, \tag{4}$$

which provides the effective rate-fidelity trade-off in elementary link generation.

III. Assumptions and Formulation

To describe the problem formally, we make the following assumptions, many of which are similar to that of [6], [7].

A1 *Static network:* Analogous to classical NUM and QNUM, we aim to distribute communication resources in a static setting. Specifically, routes and their corresponding rate-fidelity allocations are determined before the network goes into operation. We assume that each route supports a single application throughout the operational period, implying the *derived value* of an allocation remains the same for a demand over time. The network topology is also assumed to be fixed.

A2 *Entanglement swapping:* End-to-end entanglement between the terminal nodes is established via a two-step process. First, entanglement is generated at the link level between adjacent nodes. Then, entanglement swapping is performed at intermediate repeater nodes along the path to produce e2e entanglement [27]. Following [6], [7], we assume the simplified setting where link-level entanglements are generated simultaneously and swapped immediately, avoiding decoherence. Since swapping two Werner states yields another Werner state whose parameter equals the product of the initial parameters [28], this enables a concise representation of the e2e link.

A3 *Rate and fidelity of entanglement generation:* As described in Sect. II-B, the rate-fidelity trade-off of the elementary link generation is inspired by the single-click protocol and follows the realtion (4). That is, if we fix the fidelity of link $j \in [l]$ by setting its Werner parameter to w_j, the maximum entanglement generation rate will be $\mu_j := d_j(1-w_j)$, where $d_j := 3\kappa_j\eta_j/2T_j$ (4). Note that A1 implies that w_j's are set in advance and remain fixed throughout network operation. Consequently, the contributions of the jth link towards the e2e Werner parameters are identical for all routes traversing it.

A4 *Utility of a demand:* We first define the overall (binary) link-route incidence matrix A, where $a_{jm} := ((A))_{jm} = 1$ iff the mth route traverses link j. Recall that we restrict our choice of routes to simple paths. The link-route incidence matrices corresponding to valid single-path routings of the k demands belong to the following set

$$\mathcal{P}(A) := \{\tilde{A} \in \{0,1\}^{l\times k} : \tilde{A}_{:i} = A_{:m} \text{ for some } m \in I_i\} \tag{5}$$

Since each demand is served via a single route, we denote the rate allocated to demand i by x_i and the e2e Werner parameter by u_i. As the total rate allocated to demands cannot exceed a link's maximum entanglement generation rate for any valid single-path routing $\tilde{A}$, we must have $\sum_{i\in[k]} \tilde{a}_{ji}x_i \le \mu_j$, where $\tilde{a}_{ji} = ((\tilde{A}))_{ji}$. Recall from A2 that the e2e Werner parameter is the product of link-level Werner parameters. Given a routing $\tilde{A}$, we then have $u_i = \prod_{j\in[l]} w_j^{\tilde{a}_{ji}}$. We quantify the suitability of the fidelity (e2e Werner parameter to be precise) allocation u_i using a nonnegative non-decreasing function f_i, where

$$f_i : [0,1] \to [0, \beta_i]$$
$$u_m \mapsto f_i(u_m)$$

In the QNUM framework, f_i's are generally taken to be entanglement measures [22], secret key fraction or fidelity of teleportation. Following [6], the utility of a demand is assumed to have the form $x_i f_i(u_i)$. Importantly, the demand utility is a function of the routing $\tilde{A}$.

A5 *Network utility:* For a given single-path routing $\tilde{A}$, the network utility $\mathcal{U}(\tilde{A})$ is defined as the product of the demand utilities:

$$\mathcal{U}(\tilde{A}) := \prod_{i\in[k]} x_i f_i\Big(\prod_{j\in[l]} w_j^{\tilde{a}_{ji}}\Big). \tag{6}$$

Note that alternative forms for route, demand and network utility functions are possible. However, the product form guarantees that a given network utility level is reached only when each demand receives sufficient rate and fidelity. In canonical NUM, the utility is typically written as $\ln\mathcal{U} = \sum_i \ln U_i$, where U_i is the utility of demand i. In contrast, we follow the welfare economics convention [29], from which NUM originates, to motivate our formulation. Both approaches are clearly equivalent for the purpose of utility maximisation.

Based on these assumptions, we now formulate the utility-based entanglement routing problem.

A. The Utility-based Entanglement Routing Problem

We denote the rate allocation vector for the demands by $\vec{x} = (x_1, x_2, \ldots, x_k)$ and the Werner parameter vector for the links by $\vec{w} = (w_1, w_2, \ldots, w_l)$. Also, let $\tilde{A}_j$ denote the jth row of routing $\tilde{A}$. The single-path entanglement routing problem can then be written as

$$\begin{aligned}
\max_{\vec{x},\vec{w},\tilde{A}} \quad & \prod_{i\in[k]} x_i f_i\Big(\prod_{j=1}^{l} w_j^{\tilde{a}_{ji}}\Big)\Big) \\
\text{s.t.} \quad & \vec{0} \preceq \vec{x}, \text{ (Non-negative rates)} \\
& \vec{0} \preceq \vec{w} \preceq \vec{1}, \text{ (Fidelity bounds)} \\
& \langle \tilde{A}_j, \vec{x}\rangle \le \mu_j = d_j(1-w_j) \ \ \forall j \in [l], \text{ (Rate constraints)} \\
& \tilde{A} \in \mathcal{P}(A). \text{ (Single path constraint)}
\end{aligned} \tag{7}$$

Here, $\preceq$ denotes element-wise inequality and $\langle \tilde{A}_j, \vec{x}\rangle$ denotes the dot product of $\tilde{A}_j$ and $\vec{x}$. As argued in [7], the monotonicity assumption on f_i's lets us replace the inequality in the rate constraints in (7) by the equality $w_j = 1 - \langle \tilde{A}_j, \vec{x}\rangle / d_j$ and thereby eliminate $\vec{w}$:

$$\max_{\vec{x},\tilde{A}} \quad \prod_{i\in[k]} x_i f_i\Big(\prod_{j=1}^{l}\Big(1 - \frac{\langle \tilde{A}_j, \vec{x}\rangle}{d_j}\Big)^{\tilde{a}_{ji}}\Big) \tag{8}$$

$$\text{s.t.} \quad \vec{0} \preceq \vec{x}\ , \tag{9}$$

$$\langle \tilde{A}_j, \vec{x}\rangle \le d_j \quad \forall j \in [l]\ , \tag{10}$$

$$\tilde{A} \in \mathcal{P}(A)\ . \tag{11}$$

Remark 1. *Note that on a non-simple path, removal of the loops increases the value of the e2e Werner parameter and, by monotonicity of entanglement measures f_i, the value of the objective function* (8) *as well. Thus, the restricting A to simple paths can be done without loss of generality.*

Remark 2. *Entanglement measures require $f_i(u) = 0$ for $u \le 1/3$ as Werner states are separable until the threshold Werner paramter value of $1/3$. However, if we only require f_i's to be non-decreasing and non-negative, for the special case $f_i(u) = 1$ and $d_j = 1\,\forall i, j$, the objective function in* (8) *attains the value 1 iff there exist edge-disjoint paths (EDP) for the given demands. Since the EDP problem is NP-hard [30], this implies that under this relaxed assumption, the single-path entanglement routing problem cannot be solved efficiently.*

For a given demand, the number of paths is in general exponential in the number of nodes n (i.e., $|V|$), which

implies that the number of candidate routings $\tilde{A}$ in (11) grows exponentially with n. To address this, we present a link-based formulation of (8)–(11). For certain entanglement measures f_i, the problem reduces to an MICP with $6kl+k+l$ variables (of which $2kl$ are binary), or can be *closely* approximated by it.

IV. Link-based Formulation

A. Direct Formulation

To formulate the optimisation problem solely in terms of demand-based link-level variables, we first convert the undirected network graph into a directed one by replacing each undirected edge with two directed edges, which are then re-labelled. We denote the indices of the directed edges (communication links) in the new graph by j', in contrast to j in the undirected graph. We now introduce the variables required to define the single-path routing problem.

$\epsilon(j)$	the set comprising two directed link indices corresponding to the undirected link j, $j \in [l]$
$\delta^+(v)$	the set of indices of incoming links at v, $v \in V$
$\delta^-(v)$	the index set of outgoing links from v, $v \in V$
$x_{ij'}$	the rate allocated to demand i on the j'th directed link, $i \in [k], j' \in [2l]$

We group the rate allocation variables into the following allocation vector

$$\overleftrightarrow{x} := (x_{11}, x_{12}, \ldots, x_{1\,2l}, \ldots, x_{k1}, x_{k2}, \ldots, x_{k\,2l})\,. \tag{12}$$

Now, the rate allocated to demand i is given by the maximum allocation to this demand over links originating from its source node s_i, i.e., $\max_{j' \in \delta^-(s_i)} x_{ij'}$. Also, the (j,i)th element $\tilde{a}_{ji}$ of the routing matrix $\tilde{A}$ is 1 iff there is an allocation on this link in either direction, i.e., $\sum_{j' \in \epsilon(j)} x_{ij'} > 0$. We can thus rewrite the contribution of (undirected) link j to the e2e Werner parameter of the route serving demand i in (8) as

$$w_{ij}(\overleftrightarrow{x}) := \left(1 - \frac{\langle \tilde{A}_j, \vec{x} \rangle}{d_j}\right)^{\tilde{a}_{ji}} = 1 - \mathbb{1}_{\{\sum_{j' \in \epsilon(j)} x_{ij'} > 0\}} \sum_{i \in [k],\, j' \in \epsilon(j)} x_{ij'}/d_j\,, \tag{13}$$

where $\mathbb{1}_{\{\}}$ denotes the indicator function. This gives following link-based formulation of the single-path routing problem:

$$\max_{\overleftrightarrow{x}} \quad \prod_{i \in [k]} \Big(\max_{j' \in \delta^-(s_i)} x_{ij'} \Big) f_i \Big(\prod_{j=1}^{l} w_{ij}(\overleftrightarrow{x}) \Big) \tag{14}$$

$$\text{s.t.} \quad x_{ij'} \geq 0, \quad \forall i \in [k],\ \forall j' \in [2l] \tag{15}$$

$$\sum_{i \in [k],\, j' \in \epsilon(j)} x_{ij'} \leq d_j, \quad \forall j \in [l] \tag{16}$$

$$\sum_{j' \in \delta^+(v)} \mathbb{1}_{\{x_{ij'} > 0\}} \leq 1,\ \forall v \in V, \quad \sum_{j' \in \delta^+(s_i)} \mathbb{1}_{\{x_{ij'} > 0\}} = 0;\ \forall i \in [k] \tag{17}$$

$$\sum_{j' \in \delta^-(v)} \mathbb{1}_{\{x_{ij'} > 0\}} \leq 1,\ \forall v \in V, \quad \sum_{j' \in \delta^-(t_i)} \mathbb{1}_{\{x_{ij'} > 0\}} = 0;\ \forall i \in [k] \tag{18}$$

$$\sum_{j' \in \delta^+(v)} x_{ij'} - \sum_{j' \in \delta^-(v)} x_{ij'} = 0,\ \forall v \notin \{s_i, t_i\}, \forall i \in [k] \tag{19}$$

We now explain the constraints:

- *Positivity*: rate allocations must be nonnegative (15).
- *Capacity constraints*: the total bidirectional allocation on the jth (undirected) link cannot exceed d_j (16) (see (10)).
- *Single path*: for each demand, at most one incoming link (resp. outgoing) of a vertex can have a positive rate allocation (17) (resp. (18)), while there are no allocations on the incoming (resp. outgoing) links to the source (resp. destination). This also ensures absence of loops.
- *Flow preservation*: for each demand, the incoming and outgoing rates must be equal at each node, except for the corresponding source and destination (19).

Proposition 1. *The route-based* (8)–(11) (**R**) *and link-based* (14)–(19) (**L**) *formulations are equivalent.*

Proof sketch. We need to show that any feasible solution of (**R**) corresponds to a feasible solution of (**L**) with the same objective value, and vice versa. Given a feasible $(\vec{x}, \tilde{A})$ for (**R**), each demand i is assigned a unique route $\tilde{A}_{:i}$ with allocation x_i. This route decomposes into a sequence of adjacent directed links from s_i to t_i. Setting $x_{ij'} = x_i$ on these links and $x_{ij'} = 0$ elsewhere satisfies constraints (15)–(19), and the objective values match by the argument preceding (13).

Conversely, for any feasible solution of (**L**), constraint (18) implies that for each demand i, $x_{ij'}$ can be positive on at most one outgoing link from s_i. Together with flow conservation (19) and no positive outgoing flow from t_i constraint (18), this yields a unique directed simple path $\tilde{A}_{:i}$ from s_i to t_i whenever such a positive allocation exists on one outgoing link from s_i. Then x_i is set equal to the common nonzero value of $x_{ij'}$ along the route $\tilde{A}_{:i}$. If all outgoing allocations from s_i are zero, any path from s_i to t_i may be selected, yielding a corresponding feasible solution of (**R**) with zero network utility. A detailed proof will be given in an extended version. □

We now update formulation (**L**) to make it an MICP.

- *Surrogate variables for single path constraints*: we define

$$y_{ij'} = \mathbb{1}_{\{x_{ij'} > 0\}}\,,\ \forall i \in [k], \forall j' \in [2l]\,,$$

and accordingly introduce the following constraint:

$$x_{ij'} \leq d_j y_{ij'},\ x_{ij'} \geq \varepsilon y_{ij'},\ y_{ij'} \in \{0,1\},\ \forall i,\ \forall j', \tag{20}$$

where $\varepsilon > 0$ is arbitrarily small. Since the utility is zero when any demand lacks a positive allocation on at least one outgoing link from its source and one incoming link to its destination, we eliminate such cases by updating the single-path constraints (17)–(18) to:

$$\begin{gathered}\sum_{j' \in \delta^+(v)} y_{ij'} \leq 1,\ \forall v \in V \setminus \{s_i\}, \quad \sum_{j' \in \delta^+(s_i)} y_{ij'} = 0;\ \forall i \in [k] \\ \sum_{j' \in \delta^-(v)} y_{ij'} \leq 1,\ \forall v \in V \setminus \{t_i\}, \quad \sum_{j' \in \delta^-(t_i)} y_{ij'} = 0;\ \forall i \in [k] \\ \sum_{j' \in \delta^-(s_i)} y_{ij'} = 1, \quad \sum_{j' \in \delta^+(t_i)} y_{ij'} = 1;\ \forall i \in [k]\end{gathered} \tag{21}$$

Also, absence of loops implies

$$y^+_{ij} := \sum_{j' \in \epsilon(j)} y_{ij'} \leq 1,\ \text{i.e., } y^+_{ij} \in \{0,1\}\,\forall i, j. \tag{22}$$

We group the binary surrogates into vectors $\vec{y}$ and $\vec{y}_+$.

- *Surrogate variables for Werner parameters*: we define

$$\sigma_j := \sum_{j' \in \epsilon(j),\, i \in [k]} x_{ij'}/d_j \,,\ \forall j \in [l] \tag{23}$$

$$\gamma_{ij} := y^+_{ij}\, \sigma_j \,,\ \forall i \in [k] \,,\ \forall j \in [l] \tag{24}$$

$$v_{ij} := \ln(1-\gamma_{ij}) \,,\ \forall i \in [k] \,,\ \forall j \in [l] \tag{25}$$

$$z_i := \sum_{j=1}^{l} v_{ij} \,,\ \forall i \in [k] \,. \tag{26}$$

That is, the Werner parameter $w_{ij}(\overleftrightarrow{x})$ from (13) is reformulated as $e^{v_{ij}}$. Eq. (23) and (26) are linear constraints. We modify (25) to

$$v_{ij} \le \ln(1-\gamma_{ij}) \,,\ \forall i \in [k] \,,\ \forall j \in [l] \tag{27}$$

which corresponds to a convex region due to concavity of the RHS. This does not relax the problem, as the objective function is non-decreasing in v_{ij}'s. Further, using $0 \le \sigma_j \le 1$ (capacity constraint) and (22), we enforce (24) via its exact McCormick relaxation [31]:

$$\left.\begin{aligned} &\gamma_{ij} \le y^+_{ij},\ \gamma_{ij} \ge 0 \,, \\ &\gamma_{ij} \le \sigma_{ij},\ \gamma_{ij} \ge \sigma_{ij} - (1 - y^+_{ij}) \end{aligned}\right\} \ \forall i \in [k],\ \forall j \in [l] \tag{28}$$

We then replace y^+_{ij} in (28) by $\sum_{j' \in \epsilon(j)} y_{ij'}$ and denote the respective vectors of the surrogates as $\vec{\sigma}, \vec{\gamma}, \vec{v}, \vec{z}$.

- *Objective function*: instead of maximising the objective function from (14), we take logarithm and minimise its negation. Using the convention $\ln(0) = -\infty$, which acts as a barrier, the reformulated objective becomes

$$-\sum_{i \in [k]} \left(\ln\left(\max_{j' \in \delta^-(s_i)} x_{ij'} \right) + \ln f_i(e^{z_i}) \right) .$$

Now, the term $-\ln\big(\max_{j' \in \delta^-(s_i)} x_{ij'}\big)$ can be convexly reformulated using the perspective transformation [32]:

$$\lambda_i = \sum_{j' \in \delta^-(s_i)} t_{ij'} \quad \text{with} \tag{29}$$

$$t_{ij'} = \begin{cases} -y_{ij'} \ln(x_{ij'}/y_{ij'}) & y_{ij'} > 0 \\ 0 & x_{ij'} = 0,\ y_{ij'} = 0 \\ \infty & \text{otherwise,} \end{cases} \tag{30}$$

due to the presence of the constraint

$$\sum_{j' \in \delta^-(s_i)} y_{ij'} = 1 \,, \quad y_{ij'} \in \{0,1\}\ \forall i \,.$$

Here, convexity of $-\ln$ ensures that the perspective reformulation is convex in $(\overleftrightarrow{x}, \vec{y})$ [33, Sect. 3.2.6].

For the fidelity component of the objective function, we introduce the shorthand $F_i(z) := \ln f_i(e^z)$, $i \in [k]$. Following [6], from now on we only consider three entanglement measures: SKF, a lower bound to DE and negativity. The function F_i is concave for negativity, whereas for the first two measures, we use the corresponding concave envelope $\hat{F}_i$, which closely approximates F_i. For the derivation and justification of this approximation see (63) and Fig. 2 in the Appendix. We will also empirically validate the approximation accuracy in Sect. V.

The observations above imply that the single-path entanglement routing problem can be expressed as the following MICP for the specified choices of entanglement measures.

$$\min_{\substack{\overleftrightarrow{x},\, \vec{y},\, \vec{\sigma},\, \vec{\gamma},\, \vec{v},\, \vec{z}}} \sum_{i \in [k]} \Big(\lambda_i - \hat{F}_i(z_i) \Big) \tag{31}$$

$$\text{s.t.} \quad \vec{0} \preceq \overleftrightarrow{x} \,, \tag{32}$$

$$\sum_{j' \in \epsilon(j),\, i \in [k]} x_{ij'} \le d_j,\ \forall j \in [l] \tag{33}$$

$$\sum_{j' \in \delta^+(v)} x_{ij'} = \sum_{j' \in \delta^-(v)} x_{ij'},\ \forall v \notin \{s_i, t_i\},\ \forall i \in [k] \tag{34}$$

$$y_{ij'} \in \{0,1\},\ \forall j' \in [2l],\ \forall i \in [k] \tag{35}$$

$$x_{ij'} \le d_j y_{ij'},\ \epsilon(j) \ni j' ;\ \forall j' \in [2l],\ \forall i \in [k] \tag{36}$$

$$x_{ij'} \ge \varepsilon y_{ij'},\ \forall j' \in [2l],\ \forall i \in [k] \tag{37}$$

$$\sum_{j' \in \delta^+(v)} y_{ij'} \le 1,\ \forall v \in V \setminus \{s_i\},\ \forall i \in [k] \tag{38}$$

$$\sum_{j' \in \delta^-(v)} y_{ij'} \le 1,\ \forall v \in V \setminus \{t_i\},\ \forall i \in [k] \tag{39}$$

$$\sum_{j' \in \delta^-(s_i)} y_{ij'} = 1,\ \sum_{j' \in \delta^+(s_i)} y_{ij'} = 0;\ \forall i \in [k] \tag{40}$$

$$\sum_{j' \in \delta^-(t_i)} y_{ij'} = 0,\ \sum_{j' \in \delta^+(t_i)} y_{ij'} = 1;\ \forall i \in [k] \tag{41}$$

$$\sigma_j = \sum_{j' \in \epsilon(j),\, i \in [k]} x_{ij'}/d_j,\ \forall j \in [l] \tag{42}$$

$$\gamma_{ij} \le \sum_{j' \in \epsilon(j)} y_{ij'},\ \forall j \in [l] \,,\ \forall i \in [k] \tag{43}$$

$$\gamma_{ij} \le \sigma_j,\ \forall j \in [l] \,,\ \forall i \in [k] \tag{44}$$

$$\gamma_{ij} \ge 0,\ \forall j \in [l] \,,\ \forall i \in [k] \tag{45}$$

$$\gamma_{ij} \ge \sigma_j - 1 + \sum_{j' \in \epsilon(j)} y_{ij'},\ \forall j \in [l] \,,\ \forall i \in [k] \tag{46}$$

$$v_{ij} \le \ln(1-\gamma_{ij}),\ \forall j \in [l] \,,\ \forall i \in [k] \tag{47}$$

$$z_i = \sum_{j=1}^{l} v_{ij},\ \forall i \in [k] \,. \tag{48}$$

Remark 3. *The only potential source of inexactness in the formulation is the overestimator $\hat{F}_i$. For networks with sufficiently high entanglement generation rates per demand, the MICP remains exact because $\hat{F}_i$ differs from F_i only beyond a threshold $\hat{z}$. In high-rate regimes, the optimal solution does not exceed this threshold, as increasing the rate is more advantageous than improving fidelity ($f_i \le 1$ for considered entanglement measures). This can be readily verified from the solution. Otherwise, rerunning the MICP with a concave underestimator $\check{F}_i$ in place of $\hat{F}_i$ provides a bound on the approximation error, given by the gap between the two optimal values. See Fig. 2 and the appendix for further details.*

The MICP can be solved using standard MICP solvers, which remains *practical* for instances with moderate size. For larger networks, we propose an upper bound and a randomised heuristic as follows.

Upper bound: we obtain a convex relaxation of the MICP by modifying the integrality constraint (35) to: $0 \le y_{ij'} \le 1,\ \forall i, j'$. We also explicitly include the following constraints, which were automatically true earlier due to (35). The first constraint

is introduced to mitigate path splitting and it allows us to drop either (38) or (39), while the second one enforces (22).

$$\sum_{j'\in\delta^+(v)} y_{ij'} = \sum_{j'\in\delta^-(v)} y_{ij'},\ \forall v\notin\{s_i,t_i\},\ \forall i\in[k]\,, \tag{49}$$

$$\sum_{j'\in\epsilon(j)} y_{ij'} \le 1,\ \forall j\in[l]\,,\ \forall i\in[k]\,. \tag{50}$$

The relaxation provides an upper bound to the maximum achievable utility under single-path routing.

Randomised heuristic: Following the idea of randomised rounding [23], we introduce a heuristic as follows:

- *Route sampling*: for each demand, the output variables $y_{ij'}$ from the relaxed MICP are interpreted as probabilities. Employing the path-stripping algorithm [23], we extract directed paths between each source and destination, with corresponding weights. Because the total outgoing and incoming probabilities at the source and destination respectively sum to one (constraints (40) and (41)) and flow is conserved at intermediate nodes (49), this procedure indeed yields a valid probability distribution over output routes for each demand; see [23] for details.
- *Allocation optimisation for fixed routes*: once a fixed route is sampled for each demand, we use the convex QNUM formulation [7] to compute the optimal rate allocation $\vec{x}$.

Next, we propose an alternative heuristic and upper bound inspired by *minimum-congestion routing* in classical networks [23], both of which often outperform their counterparts on evaluated examples. The use of minimum-congestion routing is motivated by the dependence of a route's e2e Werner parameter on the congestion of its constituent links (13). Both the heuristic and upper bound begin by solving an LP that provides a lower bound on maximum congestion. Overall, the heuristic is computationally faster than the previous one, while the upper bound requires solving an MICP with k binary and $8k{+}1$ total variables in the second step, substantially fewer than in the original formulation.

B. Minimum Congestion-based Heuristic and Upper Bound

Definition IV.1 (Maximum congestion)**.** *For a valid routing $\tilde{A}$ as introduced in* (5)*, its maximum congestion is defined as the highest number of routes traversing any single link, i.e.,*

$$c_{\max}(\tilde{A}) = \max(\tilde{A}\vec{1})\,.$$

As shown in [23, Sect. 3], minimising the maximum congestion over all possible routings can be formulated as a mixed-integer linear program and then relaxed to an LP for large networks. Let $\underline{c}$ denote the optimal congestion value obtained from the LP; then $\lceil\underline{c}\rceil$ provides a lower bound on the maximum congestion. By interpreting the fractional LP solution as probabilities, [23] proposes a randomised routing strategy, which forms the basis of our heuristic.

Randomised heuristic: For each demand, a route is sampled according to the randomised routing strategy of [23, Sect. 3]. Given these sampled routes, the convex QNUM formulation [7] is applied to compute the optimal rate allocation $\vec{x}$. This heuristic is computationally faster than the previous one, as the min-congestion LP is faster than the MICP's relaxation.

Upper bound: first, we introduce the following notation: the jth largest element among the link-specific constants $\vec{d} := (d_1, d_2, \ldots, d_l)$ is denoted by $d_{(j)}$. Given a lower bound to the maximum congestion level $\lceil\underline{c}\rceil$, we can obtain an upper bound to the achievable network utility by considering the following optimistic scenario:

- *Single-link congestion:* we assume all but one link support at most one route, while the remaining (congested) link experiences a congestion level of $\lceil\underline{c}\rceil$. The corresponding rate–fidelity trade-off constant is taken as $d_{(1)}$.
- *Shortest-path routing:* all demands are assumed to be routed through their respective shortest paths, with λ_i denoting the corresponding path length for demand $i \in [k]$. Furthermore, the e2e Werner parameter can be upper bounded by assuming that the rate–fidelity trade-off constants along uncongested links attain their best possible combination: $\{d_{(1)}, d_{(2)}, \ldots, d_{(\lambda_i-1)}\}$.

Determining the maximum achievable network utility then reduces to selecting the subset of demands routed through the congested link. Let $\alpha_i \in \{0,1\}$ denote the route assignment variable for demand i, and let $\vec{\alpha}$ denote the corresponding vector. The upper bound can then be formally expressed as follows. The proof follows directly from the reasoning above and is omitted due to space constraints.

Proposition 2. *Given a maximum congestion level $\lceil\underline{c}\rceil$, the maximum achievable network utility* (8) *can be bounded as*

$$\max_{\substack{\vec{0}\preceq\vec{x},\,\tilde{A}\vec{x}\preceq\vec{d}\\ \tilde{A}\in\mathcal{P}(A)}} \prod_{i\in[k]} x_i f_i\Big(\prod_{j=1}^{l}\Big(1-\frac{\langle\tilde{A}_j,\vec{x}\rangle}{d_j}\Big)^{\tilde{a}_{ji}}\Big) \tag{51}$$

$$\le \max_{\substack{0\le x_i\le d_{(\lambda_i)}\\ \vec{\alpha}'\vec{1}=\lceil\underline{c}\rceil}} \prod_{i\in[k]} x_i f_i\Big(\Big(1-\frac{x_i+\alpha_i\sum_{i'\neq i}\alpha_{i'}x_{i'}}{d_{(1)}}\Big)\prod_{j\in[\lambda_i-1]}\Big(1-\frac{x_i}{d_{(j)}}\Big)\Big).$$

The RHS of (51) is formulated as the following optimisation problem. We again minimise the negative logarithm of the objective, introduce surrogates for the e2e Werner parameters and approximate $F_i(z)(=\ln f_i(e^z))$ by $\hat{F}_i(z)$.

$$\min_{\substack{\vec{x},\vec{\alpha},\vec{\eta},S,\vec{\delta}\\ \vec{\zeta},\vec{\gamma},\vec{v},\vec{z}}} -\sum_{i\in[k]}\ln x_i - \sum_{i\in[k]}\hat{F}_i(z_i) \tag{52}$$

$$\text{s.t. } 0\le x_i\le d_{(\lambda_i)},\ \forall i\in[k] \tag{53}$$

$$\alpha_i\in\{0,1\},\ \forall i\in[k]\,,\quad \vec{\alpha}'\vec{1}=\lceil\underline{c}\rceil \tag{54}$$

$$\eta_i = \alpha_i x_i,\ \forall i\in[k] \tag{55}$$

$$S = \sum_{i\in[k]}\eta_i \tag{56}$$

$$\delta_i = S-\eta_i,\ \forall i\in[k] \tag{57}$$

$$\zeta_i = \alpha_i\delta_i,\ \forall i\in[k] \tag{58}$$

$$\gamma_i = x_i+\zeta_i,\ \forall i\in[k] \tag{59}$$

$$v_{i1} = \ln\big(1-\gamma_i/d_{(1)}\big)\,,\ \forall i\in[k] \tag{60}$$

$$v_{ij} = \ln\big(1-x_i/d_{(j-1)}\big),\ \forall j\in[\lambda_i]\backslash\{1\},\forall i\in[k] \tag{61}$$

$$z_i = \sum_{j\in[\lambda_i]} v_{ij},\ \forall i\in[k] \tag{62}$$

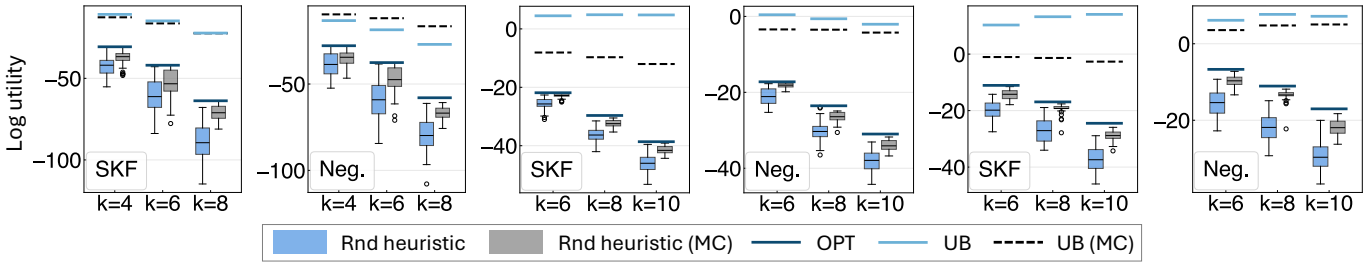

Fig. 1: Performance of the randomised heuristics (Rnd heuristic) and upper bounds (UB) vis-à-vis the optimum log utility (OPT) calculated via MICP (31)–(48), min-congestion-based metrics are marked as (MC). Results are shown for entanglement measures SKF and negativity (Neg.) on BREN (10 nodes, left 2 subplots), UNIC (15 nodes) and ARNES (17 nodes, right 2 subplots) networks [34] for varying demand counts k. The MC heuristic outperforms its counterpart *on average*, and the MC upper bound is often closer to OPT. Recall that log utility here equals utility as per NUM [4], [6].

Since $\alpha_i \in \{0,1\}$, we obtain an MICP formulation by replacing (55) and (58) with their exact McCormick relaxations:

$$\eta_i \le d_{(\lambda_i)}\alpha_i,\ \eta_i \ge 0,\ \ \eta_i \le x_i,\ \eta_i \ge x_i - d_{(\lambda_i)}(1-\alpha_i)\ \ \forall i$$
$$\zeta_i \le d_{(1)}\alpha_i,\ \zeta_i \ge 0,\ \ \zeta_i \le \delta_i,\ \ \zeta_i \ge \delta_i - d_{(1)}(1-\alpha_i)\ \ \forall i\ ,$$

and modifying (60) and (61) by making v_{ij}s less or equal to respective RHS, which keeps the problem unchanged as $\hat{F}_i$'s are increasing. In contrast to the original MICP, which involves $2kl$ binary variables, the present formulation requires only k such variables, making it *practical* to solve under wider circumstances. For instances with a large number of demands, the integrality constraints can be relaxed to obtain an upper bound.

V. Numerical Evaluations

In this section, we empirically evaluate the proposed heuristics and upper bounds on real-world networks. We consider the BREN (10 nodes, 11 links), UNIC (15 nodes, 17 links) and ARNES (17 nodes, 20 links) topologies from Topology-Bench [34], a repository of optical fibre networks. The network sizes are selected such that the optimal utility computation via the MICP (31)–(48) remains tractable, which we use for benchmarking the heuristics and upper bounds. The number of demands k is chosen proportional to the network size n, specifically $k \in \{4,6,8\}$ for BREN and $k \in \{6,8,10\}$ for others. Further, demands (SD pairs) are added *incrementally* according to the table on the right. We refer the reader to [34] for detailed network information, including topology, link lengths, node locations and name–ID mapping. Also, we show the evaluation only for two entanglement measures: SKF and negativity.

Given the link lengths L_j [34], the link-level constants d_j are computed from (4) using $\eta_j = 10^{-0.02L_j}$ and by setting $\kappa_j = 0.1$, $T_j = 10^{-3}$s for all links, which represents current state of hardware efficiency [6], [7]. Using the constants d_j and network adjacency structure, we solve the MICP (31)–(48) to obtain the optimal utility; its relaxation provides the first upper bound. The min-congestion-based upper bound is derived by solving the second MICP (52)–(62) with k binary variables. We use MOSEK solver with CVXPY for our computations. Since CVXPY only allows functions adhering to its disciplined convex programming syntax, we use piecewise linear approximation of $\hat{F}_i$ (31) with high granularity.

Both upper bounds and the corresponding randomised heuristics are compared against the optimum in Fig. 1. We observe that the min-congestion-based upper bound often outperforms its counterpart, and its heuristic also achieves higher *average* performance. While the first MICP (31)–(48) computes the maximum network utility exactly for negativity, we calculate the error of approximation for SKF by rerunning it with the concave underestimator $\check{F}_i$ in place of $\hat{F}_i$ in (31). The approximation error is not shown in Fig. 1 as the highest observed relative error was 0.00051% (ARNES, 10 demands), empirically supporting our claim of accuracy of the formulation. Finally, recall that the logarithmic utility in Fig. 1 is equivalent to the utility in NUM parlance [4], [6]; we adopted this form [29] only to motivate our formulation in Sect. III.

Network	Demands (incremental)
BREN	{(6,10),(10,4),(3,10),(9,1)}, {(5,1),(8,10)}, {(7,2),(3,7)}
UNIC	{(8,11),(5,13),(9,1),(2,8),(11,1),(10,2)}, {(7,3),(11,14)}, {(15,14),(1,9)}
ARNES	{(2,3),(10,5),(16,2),(12,7),(3,4),(17,2)}, {(15,12),(3,11)}, {(3,15),(2,4)}

VI. Conclusion

In this work, we considered the single-path version of the utility-based entanglement routing problem, aiming to determine optimal routes that maximise overall network utility. Network utility was quantified using SKF, a lower bound to DE and negativity as entanglement measures. We formulated the problem as an MICP, which provides exact solutions when negativity is used or when the network supports sufficiently high entanglement generation rates. In other cases, the formulation yields a reasonably close approximation, as seen in real-world examples. To ensure scalability in large networks, we proposed a randomised rounding-based heuristic and an upper bound derived from the relaxed MICP. We further proposed a computationally faster randomised heuristic and an upper bound based on min-congestion routing, which often outperform their counterparts on real-world network topologies. Our framework can be used for extending classical flow-based and QoS-aware routing principles to quantum networks, enabling fair and efficient allocation of quantum communication resources.

References

[1] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing", in *Proceedings of the IEEE International Conference on Computer Systems and Signal Processing*, 1984.

[2] V. Giovannetti, S. Lloyd, and L. Maccone, "Quantum-enhanced measurements: beating the standard quantum limit", in *Science*, vol. 306, no. 5700, pp. 1330–1336, 2004.

[3] A. Broadbent, J. Fitzsimons, and E. Kashefi, "Universal blind quantum computation", in *2009 50th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, 2009, pp. 517–526.

[4] F. Kelly, "Charging and rate control for elastic traffic", in *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[5] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability", in *Journal of the Operational Research Society*, vol. 49, no. 3, 1998.

[6] G. Vardoyan and S. Wehner, "Quantum network utility maximization", in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, vol. 1, pp. 1238–1248, 2023.

[7] S. Kar and S. Wehner, "Convexification of the Quantum Network Utility Maximisation Problem", in *IEEE Transactions on Quantum Engineering*, 2024.

[8] G. Apostolopoulos, R. Guerin, S. Kamat, and S. K. Tripathi, "Quality of service based routing: A performance perspective," *Proceedings of the ACM SIGCOMM*, 1998.

[9] X. Lin and N. B. Shroff, "The multi-path utility maximization problem," in *Proceedings of the Annual Allerton Conference on Communication, Control, and Computing*, vol. 41, no. 2, pp. 789–798, 2003.

[10] J. Wu, M. Lu, and F. Li, "Utility-based opportunistic routing in multi-hop wireless networks," in *Proceedings of the 28th International Conference on Distributed Computing Systems (ICDCS)*, pp. 470–477, 2008.

[11] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[12] M. Charikar, Y. Naamad, J. Rexford, and X. K. Zou, "Multi-commodity flow with in-network processing," *International Symposium on Algorithmic Aspects of Cloud Computing*, pp. 73–101, 2018.

[13] X. Lin and N. B. Shroff, "Utility maximization for communication networks with multipath routing," *IEEE Transactions on Automatic Control*, vol. 51, no. 5, pp. 766–781, 2006.

[14] A. Abane, M. Cubeddu, V. S. Mai, and A. Battou, "Entanglement routing in quantum networks: A comprehensive survey," in *IEEE Transactions on Quantum Engineering*, 2025.

[15] R. Van Meter, T. Satoh, T. D. Ladd, W. J. Munro, and K. Nemoto, "Path selection for quantum repeater networks," in *Networking Science*, vol. 3, no. 1, pp. 82–95, 2013.

[16] L. Gyongyosi and S. Imre, "Entanglement-gradient routing for quantum networks," in *Scientific Reports*, vol. 7, no. 1, p. 14255, 2017.

[17] M. Pant, H. Krovi, D. Towsley, L. Tassiulas, L. Jiang, P. Basu, D. Englund, and S. Guha, "Routing entanglement in the quantum internet," in *npj Quantum Information*, vol. 5, no. 1, p. 25, 2019.

[18] K. Chakraborty, D. Elkouss, B. Rijsman, and S. Wehner, "Entanglement distribution in a quantum network: A multicommodity flow-based approach," in *IEEE Transactions on Quantum Engineering*, vol. 1, 2020.

[19] J. Li, M. Wang, K. Xue, R. Li, N. Yu, Q. Sun, and J. Lu, "Fidelity-guaranteed entanglement routing in quantum networks," in *IEEE Transactions on Communications*, vol. 70, no. 10, pp. 6748–6763, 2022.

[20] S. Pouryousef, N. K. Panigrahy, and D. Towsley, "A quantum overlay network for efficient entanglement distribution," in *Proceedings of IEEE INFOCOM– IEEE Conference on Computer Communications*, 2023.

[21] L. Le and T. N. Nguyen, "DQRA: Deep quantum routing agent for entanglement routing in quantum networks," in *IEEE Transactions on Quantum Engineering*, vol. 3, pp. 1–12, 2022.

[22] M. B. Plenio and S. Virmani, "An introduction to entanglement measures", in *Quantum Inf. Comput.*, vol. 7, no. 1, pp. 1–51, 2007.

[23] P. Raghavan and C. D. Tompson, "Randomized rounding: a technique for provably good algorithms and algorithmic proofs", in *Combinatorica*, vol. 7, no. 4, pp. 365–374, 1987.

[24] C. Cabrillo, J. I. Cirac, P. Garcia-Fernandez, and P. Zoller, "Creation of entangled states of distant atoms by interference", in *Physical Review A*, vol. 59, no. 2, pp. 1025, 1999.

[25] W. Dür and H. J. Briegel, "Entanglement purification and quantum error correction", in *Reports on Progress in Physics*, vol. 70, no. 8, 2007.

[26] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, "Mixed-state entanglement and quantum error correction," in *Physical Review A*, vol. 54, no. 5, pp. 3824, 1996.

[27] J.-W. Pan, D. Bouwmeester, H. Weinfurter, and A. Zeilinger, "Experimental entanglement swapping: entangling photons that never interacted", in *Physical Review Letters*, vol. 80, no. 18, p. 3891, 1998.

[28] W. J. Munro, K. Azuma, K. Tamaki, and K. Nemoto, "Inside quantum repeaters", in *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 21, no. 3, pp. 78–90, 2015.

[29] P. O. Johansson, "An Introduction to Modern Welfare Economics", *Cambridge University Press*, 1991.

[30] V. Guruswami, S. Khanna, R. Rajaraman et. al., "Near-optimal hardness results and approximation algorithms for edge-disjoint paths and related problems," *Proceedings of the STOC*, pp. 19–28, 1999.

[31] G. P. McCormick, "Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems," Mathematical Programming, vol. 10, no. 1, pp. 147–175, 1976.

[32] O. Günlük and J. Linderoth, "Perspective reformulations of mixed integer nonlinear programs with indicator variables," Mathematical Programming, vol. 124, no. 1, pp. 183–205, 2010.

[33] S. Boyd and L. Vandenberghe, "Convex optimization", Cambridge University Press, 2004.

[34] R. Matzner, A. Ahuja, R. Sadeghi et. al., "Topology Bench: systematic graph-based benchmarking for core optical networks," *Journal of Optical Communications and Networking*, vol. 17, no. 1, pp. 7–27, 2024.

APPENDIX

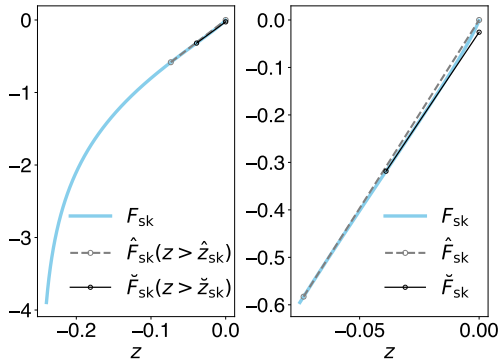

Fig. 2: The estimators of F_i when f_i=SKF (left) with zoomed in on domain of non-concavity of F_i (right). Error bounds: $\|\hat{F}_{\text{sk}} - F_{\text{sk}}\| < 0.0165$, $\|F_{\text{sk}} - \check{F}_{\text{sk}}\| < 0.0258$,

Due to space limitations, we refer the reader to [7, Eq. 16-18] for definitions of the considered entanglement measures: SKF, a lower bound to DE (as adopted in [6], [7]) and negativity in terms of the Werner parameter. Recall that for each demand i, we transform the entanglement measure f_i to $F_i(z) = \ln(f_i(e^z)), z \in (z^{(i)}_{\min}, 0]$. As F_i is concave for negativity but not for the other two measures, we look for a concave overestimator $\hat{F}_i$. The key property that we use is that F_i has a unique inflection point $\check{z}_{(i)}$ and F_i is concave in $(z^{(i)}_{\min}, \check{z}_{(i)}]$.

Concave enevlope $\hat{F}_i$: We first derive the smallest linear upper bound to F_i passing through $(0, F_i(0))$. This can be found by calculating the minimal solution of $F_i'(z) = F_i(z)/z$, as $F_i(0) = 0$. The minimal solution $\hat{z}_{(i)}$ is found via the Newton-Raphson method by initialising near $z^{(i)}_{\min}$. We then define

$$\hat{F}_i(z) = \begin{cases} F_i(z)\,, & z^{(i)}_{\min} < z \leq \hat{z}_{(i)} \\ F_i(\hat{z}_{(i)}) + F_i'(\hat{z}_{(i)})(z - \hat{z}_{(i)})\,, & \hat{z}_{(i)} < z \leq 0 \end{cases} \tag{63}$$

Now, considering the epigraph of $-F_i$ and drawing its convex hull, we can see that $\hat{F}_i$ is indeed the concave envelope of F_i.

If the z values in the solution of the MICP (31)–(48) are below $\hat{z}_{(i)}$, the MICP is exact for the entanglement routing problem. Otherwise, we rerun the MICP (31)–(48) with the underestimator $\check{F}_i$ to bound the approximation error.

Underestimator $\check{F}_i$: We simply use the tangent of F_i at the inflection point $\check{z}_{(i)}$ to define $\check{F}_i$ beyond $\check{z}_{(i)}$:

$$\check{F}_i(z) = \begin{cases} F_i(z)\,, & z^{(i)}_{\min} < z \leq \check{z}_{(i)} \\ F_i(\check{z}_{(i)}) + F_i'(\check{z}_{(i)})(z - \check{z}_{(i)})\,, & \check{z}_{(i)} < z \leq 0 \end{cases} \tag{64}$$

We show the accuracy of estimation by plotting the linear parts of $\hat{F}_i$ and $\breve{F}_i$ vis-à-vis F_i in Fig. 2 for SKF. The plots for DE are qualitatively similar with $\|\hat{F}_{\text{de}} - F_{\text{de}}\| < 0.0135$, $\|F_{\text{de}} - \breve{F}_{\text{de}}\| < 0.0212$.