Quantum Network Utility Maximization

Vardoyan, Gayane; Wehner, Stephanie

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Quantum Network Utility Maximization

Gayane Vardoyan*‡ and Stephanie Wehner*†‡

*QuTech, Delft University of Technology

†Kavli Institute of Nanoscience, Delft University of Technology

‡Quantum Computer Science, Electrical Engineering, Mathematics and Computer Science, Delft University of Technology

Abstract—Network Utility Maximization (NUM) is a mathematical framework that has endowed researchers with powerful methods for designing and analyzing classical communication protocols. NUM has also enabled the development of distributed algorithms for solving the resource allocation problem, while at the same time providing certain guarantees, *e.g.*, that of fair treatment, to the users of a network. We extend here the notion of NUM to quantum networks, and propose three quantum utility functions – each incorporating a different entanglement measure. We aim both to gain an understanding of some of the ways in which quantum users may perceive utility, as well as to explore structured and theoretically-motivated methods of simultaneously servicing multiple users in distributed quantum systems. Using our quantum NUM constructions, we develop an optimization framework for networks that use the single-photon scheme for entanglement generation, which enables us to solve the resource allocation problem while exploring rate-fidelity tradeoffs within the network topologies that we consider. We learn that two of our utility functions, which are based on distillable entanglement and secret key fraction, are in close agreement with each other and produce similar solutions to the optimization problems we study. While these two utilities place a higher emphasis on end-to-end fidelity, our third utility– based on entanglement negativity – has more favorable mathematical properties, and tends to place a higher value on the rate at which users receive entangled resources. These contrasting behaviors thus provide ideas regarding the suitability of quantum network utility definitions to different quantum applications.

Index Terms—entanglement distribution, network utility maximization, quantum network, resource allocation

I. INTRODUCTION

Quantum networks enable a host of applications whose benefits are impossible to glean with classical means alone. Notable examples of such capabilities are provably secure communication [1], [2], improved sensing [3], [4], and blind quantum computation (BQC) [5], [6]. To support such applications, quantum networks are expected to be able to produce and distribute entangled states of sufficiently high quality to nodes that request them. Efficient entanglement distribution is thus a task that is central to quantum communication, and one that is especially consequential during the noisy intermediate-scale quantum (NISQ) era [7]. Further, settings where multiple user groups demand quantum resources from a network pose a challenge not only in determining optimal entanglement distribution algorithms, but also in the very definition and interpretation of optimality – whether from the perspective of a single user group or from that of the entire network.

To address the need for quantifying a network's ability to serve user needs, we present here the notion of *quantum network utility*, which serves as an analogue to the classical concept of network utility. The latter framework was first proposed in the seminal work of Kelly *et al.* [8], [9] as a means of defining utility in classical communication networks, and subsequently being able to optimally (or fairly) allocate resources so as to maximize the sum of user or application utilities – a process that is known as Network Utility Maximization (NUM). NUM's introduction to the networking community spurred a large body of research that resulted in both a rich theory as well as a family of algorithms for distributed resource allocation, some of which were even commercialized [10].

By adapting the classical NUM framework to quantum networks, we aim to enable the quantum networking community to model and analyze distributed quantum architectures within a formal mathematical framework, potentially leading to the design of novel distributed resource allocation protocols that can service the needs of all communicating parties in a network. In extending the notion of NUM to quantum systems, however, one is inevitably confronted with a number of uniquely quantum problems:

(1) *Which methods can be used to cope with resource interdependence in quantum networks?* In classical networks, physical channels serve as conduits through which data flows from a set of sources to a set of destinations. In contrast to the classical case, however, there is no notion of "traffic" in networks that distribute entanglement by first generating "elementary" entangled links between neighboring nodes and later converting these into longer-distance links via entanglement swapping [11], [12] (*i.e.*, quantum networks made up of first- and second-generation quantum repeaters [13]), see Figure 1 for an illustration. The network instead consists of independently-operating physical links that probabilistically produce resources – entanglement – which then, after undergoing a collective transformation, are consumed by groups of users. These state transformations (*e.g.*, entanglement swapping, distillation) result in interdependence between the nodes, links, and the states that they generate.

(2) *How does a quantum user (an end node that runs a quantum application) derive utility from the network?* As with classical network utility, the precise answer to this question has a dependence on the application under consideration. In a classical network, an individual user's utility may be expressed in terms of rate, delay, jitter,
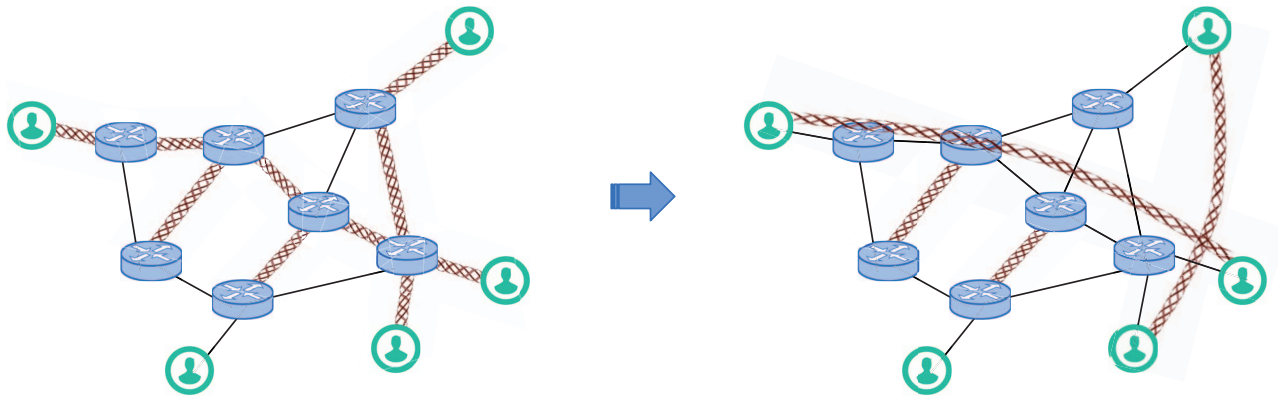
Fig. 1: A quantum network consisting of a number of users and quantum repeaters. After links successfully generate entanglement (left), repeaters perform entanglement swapping to distribute end-to-end states to users (right).

and other perceptions of quality of service (QoS). An e-mail sender, for example has different requirements and expectations than a video watcher. In the same way, users who, for example, execute quantum key distribution (QKD) have a different understanding of QoS than a user who runs BQC. However, in a quantum network, since entanglement is consumed by groups of users, an individual user's utility is always coupled to at least one other user's utility, and this joint utility in turn has a dependence both on the rate at which entanglement is allotted, as well as on the quality of each state. A utility function should then capture both classical (*e.g.*, rate) as well as quantum (*e.g.*, fidelity) QoS measures.

(3) *How should rate-fidelity tradeoffs, which are inherent in quantum networks, be captured?* In a classical network, a source may suffer diminishing returns as its sending rate is increased beyond a threshold, but it does not typically experience a *decrease* in utility; *i.e.*, traditionally, utility functions are concave and non-decreasing. In a quantum network, in contrast, one pays a cost for generating states with higher fidelity – whether this is done via entanglement distillation or through other means (*e.g.*, by adjusting the bright-state population when performing the single-photon entanglement generation protocol [14], [15]) – a decrease in the rate at which such states are successfully generated. A consequence of this is that utility is not necessarily a non-decreasing function of classical *or* quantum QoS measures, and may in fact approach 0 (or $-\infty$, depending on the range of the utility function) in cases where performance favors one measure exceedingly more than the others.

We make the following contributions towards answering these questions:

- We define and study three quantum network utility functions, each based on a different entanglement measure: distillable entanglement [16], [17], secret key rate [18], [19], and entanglement negativity [20];
- We develop a quantum NUM (QNUM) framework based

on the single-photon entanglement generation scheme;
- We apply the framework to a number of network topologies and solve the resource allocation problem by optimizing the rate-fidelity tradeoff on each physical link in the network.

We focus on applications that require bipartite entanglement generation between end users, to which we also refer to as end-to-end bipartite entanglement, as this is the most basic form of entanglement that is nevertheless a vital component of a diverse set of distributed quantum applications. To capture rate-fidelity tradeoffs as discussed in (3) above, we use the single-photon entanglement generation scheme, wherein the bright-state population parameter directly affects both the quality of states and the rate at which they are generated (see more in Sections II and III). Entanglement distillation, while a necessary process for long-distance entanglement distribution in near-term networks, is left as a subject of follow-up work.

The remainder of this manuscript is organized as follows: in Section II, we provide relevant background on classical NUM. In Section III, we state our assumptions and define the problem in detail. In Section IV, we define three QNUM formulations and apply them to different network scenarios in Section V. We conclude in Section VI.

II. BACKGROUND

NUM in its canonical form is the optimization problem

$$\max_{x_r} \sum_{r \in \mathcal{R}} U_r(x_r), \tag{1}$$

subject to

$$\sum_{r:l \in r} x_r \leq c_l, \ \forall l \in \mathcal{L}, \tag{2}$$

$$x_r \geq 0, \ \forall r \in \mathcal{R}, \tag{3}$$

where x_r corresponds to the rate at which a user sends packets across route r, c_l is link l's capacity, and U_r is a utility function [21]. Above, $\mathcal{L}$ and $\mathcal{R}$ are sets containing all links and routes in the network, respectively. U_r is usually assumed to be strictly

concave and may be used in one of two ways: either to describe the way in which a user (typically a source sending data) derives satisfaction from the network, or as a function that is *assigned* to a user by the network to control the manner in which resources are allocated. A popular example of the latter is the idea of fair treatment, specifically, functions of the form

$$U_r(x_r) = \frac{x_r^{1-\alpha}}{1-\alpha}, \qquad (4)$$

for $\alpha > 0$, are known as α-fair, each enforcing a different notion of fairness during resource allocation. In this work, we will make use of the so-called proportionally-fair utility function $\log(x_r)$, which is obtained by letting $\alpha \to 1$ in (4).

The classical NUM formulation stated above can be used as a starting point for developing distributed resource allocation algorithms which can operate in the presence of feedback delays and changing network conditions. Following its development in the late nineties, NUM has proven itself highly influential in the design of distributed classical network algorithms (*e.g.*, CSMA scheduling for multi-hop wireless networks [22], several congestion control algorithms, *e.g.*, FAST TCP [10], and utility-based resource allocation algorithms in OFDM networks [23]). While some of these algorithms have been commercialized, certain others, *e.g.*, TCP Reno and Vegas [24], have been shown to be reverse-engineerable from NUM. Given NUM's success in the classical study of distributed architectures, our aim is to investigate the adaptability of these methods to quantum applications. To our knowledge NUM has not been, to date, extended to quantum networks.

III. PROBLEM FORMULATION

Classical NUM in its basic form is most applicable under static and continual conditions: *i.e.*, there is a fixed number of user communication sessions, and traffic is injected into the network *ad infinitum*. We will make an analogous assumption for this initial QNUM formulation: namely, that all pairs of users wishing to communicate are fixed, and that each communication session has an infinite backlog of entanglement requests. In principle, there are a number of ways to accommodate changing network dynamics. In the case of a centralized controller, for instance, one may update and re-solve the optimization program after some time has elapsed; this action could be taken, *e.g.*, after a fixed/pre-determined period, whenever an existing communication session drops or a new one arrives, or according to some other heuristic. In the case of a completely distributed setting, our proposed QNUM formulation can be distributized (using, *e.g.*, the primal approach after defining suitable cost functions on the constraints of the problem [21]) so that each communication session and each link are responsible for tuning their rate and Werner parameter, respectively, according to feedback received from the network. Another, albeit often less tractable technique, is to transform the problem into a stochastic NUM formulation, see [25] for an overview.

As with basic classical NUM, we assume that routes are pre-determined and fixed for all communication sessions, as

doing otherwise would require incorporating entanglement routing – a complex problem in its own right – into the QNUM formulation. A route corresponding to a communication session between two users in a network is defined by the path connecting the users. Throughout this manuscript, we often use the terms "communication session" and "route" interchangeably (similar to the interchangeable use of "source" and "route" in classical NUM literature). The set $\mathcal{R}$ contains all active communication sessions. The set $\mathcal{L}$ represents the links of a quantum network. We will often use the notation $l \in r$ to denote that link l is a constituent of route r. An active communication session experiences QoS both in terms of the rate at which end-to-end entanglement is served to it by the network – R_r for route r, as well as the quality (fidelity) of said states.

In this work, we assume that elementary link-level entanglement is generated using the single-photon scheme. In this scheme, a state of the form

$$\rho = (1-\alpha)|\Psi^+\rangle\langle\Psi^+| + \alpha|\uparrow\uparrow\rangle\langle\uparrow\uparrow| \qquad (5)$$

is generated with probability $p_{\text{elem}} = 2\eta\alpha$, where $|\Psi^+\rangle$ is a Bell state that is orthogonal to the bright state $|\uparrow\uparrow\rangle$, η is the transmissivity between one end of the link and its midpoint (where a midpoint heralding station is located, consisting of a beamsplitter and two detectors), and α is the bright-state population – a tunable parameter that enables the rate-fidelity tradeoff [15], [26][1]. We use the following formula for the transmissivity of link l that has length L_l:

$$\eta_l = 10^{-0.1\beta L_l}, \qquad (6)$$

where $\beta = 0.2$ dB/km is the fiber attenuation coefficient. In numerical evaluation, we will often multiply p_{elem} by a constant factor $c \in (0,1)$ to account for various system inefficiencies other than the loss in fiber.

To simplify post-swap fidelity calculations, we assume that instead of the state (5), each link generates Werner states, which may be written as

$$\rho_w = w|\Psi^+\rangle\langle\Psi^+| + (1-w)\frac{\mathbb{I}_4}{4}, \qquad (7)$$

parametrized above in terms of the Werner parameter w which relates to ρ_w's fidelity as

$$F = \frac{3w+1}{4}. \qquad (8)$$

We thus assume that Werner states are generated with success probability p_{elem} and fidelity $1-\alpha$. While this approximation will result in lower end-to-end fidelities – depolarizing noise is considered a worst-case scenario – it nevertheless allows us to obtain lower-bounds on performance. Moreover, due to the simplicity of the post-swap Werner parameter w^{e2e}, given for a route r by

$$w_r^{\text{e2e}} = \prod_{l \in r} w_l, \qquad (9)$$

---

[1]Note that this description of the state ρ and probability p_{elem} applies in high-loss regimes, *i.e.*, when η is small.

where w_l are Werner parameters of elementary links involved in the swap (see appendices of [27] for a proof), we can easily construct utility functions with favorable mathematical properties. We finally assume that each link l has an associated repetition time T_l that specifies the time between entanglement generation attempts. We may thus express a link's entanglement generation rate as

$$\frac{p_{\text{elem}}}{T} = \frac{2c\eta\alpha}{T} = \frac{2c\eta(1-F)}{T} = \frac{3c\eta}{2T}(1-w) \equiv d(1-w), \tag{10}$$

where we have defined $d := 3c\eta/2T$. We emphasize that there are two different notions of rate in our QNUM formulation: the first is that of a link's entanglement generation rate, given by $d_l(1 - w_l)$, and the second is the entanglement rate R_r allotted to route r by the network. A link will subdivide its rate across all routes that it services, and the output of the optimization procedure will determine both the link's rate, as well as its apportionment across active communication sessions. We finally note that configuring the bright-state population parameter of a link is equivalent to configuring the fidelity of the resulting states or, in our approximation, the Werner parameter; we find the latter most convenient to use within the optimization framework.

IV. QUANTUM NETWORK UTILITY MAXIMIZATION FRAMEWORK

In this section we present a general QNUM framework that can accommodate single-photon entanglement generation schemes and as a result, allows one to explore rate-fidelity tradeoffs found in such quantum networks. In its most basic form, the QNUM formulation is as follows (for mathematical convenience, instead of maximizing aggregate utility, we minimize its negation):

$$\min \quad -\sum_{r \in \mathcal{R}} U_r(R_r, \boldsymbol{w}) \tag{11}$$

$$\text{subject to} \sum_{r:l \in r} R_r - d_l(1 - w_l) = 0, \quad \forall l \in \mathcal{L}, \tag{12}$$

$$0 \leq w_l \leq 1, \quad \forall l \in \mathcal{L}. \tag{13}$$

Above, $\boldsymbol{w} = \{w_1, \ldots, w_{|\mathcal{L}|}\}$ is the vector of all elementary-link Werner parameters.

One may easily incorporate minimum fidelity requirements imposed on end-to-end entangled states generated on routes, by including constraints of the form

$$\prod_{l \in r} w_l \geq \frac{4F_r^\star - 1}{3}, \quad \forall r \in \mathcal{R}, \tag{14}$$

where $F_r^\star$ denotes the threshold fidelity for route r. Note that constraint (12) is equivalent to $\sum_{r:l \in r} R_r = (2c_l\eta_l/T_l)(1-F_l)$, *i.e.*, the aggregate rate at which users/applications receive entanglement at link l is equal to link l's entanglement generation rate. In principle one could replace the equality with a less-than sign, but we opt for the former to avoid entanglement wastage resulting from a link that produces a

surplus of entanglement which then goes unused by the routes. We can convert the inequality constraints (14) to a set of convex constraints by taking the logarithm of both sides, to obtain

$$\log\left(\frac{4F_r^\star - 1}{3}\right) - \sum_{l \in r} \log(w_l) \leq 0. \tag{15}$$

If the utilities U_r are concave, then this formulation results in a convex optimization problem, as the equality constraints (12) are affine (*i.e.*, each is a sum of a linear function and a constant).

Following, we introduce three utility functions which incorporate useful measures of entanglement distribution rate and quality. While two of these measures possess operational meanings, we find that the utility functions which we define on them are not concave in general – this means that there are few theoretical guarantees for finding global optima. The third utility function on the other hand is concave, albeit it is based on an entanglement measure that is not associated with a significant operational interpretation. While in principle there are an infinite number of quantum network utility functions, we have identified these three as ones that encompass a number of important properties and serve well in identifying favorable utility function behaviors.

A. QNUM based on Distillable Entanglement

For bipartite states, distillable entanglement can be viewed as the maximum rate at which nearly-perfect Bell states can be distilled from many copies of mixed bipartite states, using local operations and classical communication. While for isotropic[2] and Werner states distillable entanglement can be computed numerically, we opt to work here with a lower-bound on this entanglement measure, given by

$$D(F) \geq \max\left(1 + F\log_2 F + (1 - F)\log_2 \frac{1 - F}{3}, 0\right), \tag{16}$$

for $F \in [1/2, 1]$, see [17]. Note that the quantity

$$D_H(F) := 1 + F\log_2 F + (1 - F)\log_2 \frac{1 - F}{3}, \tag{17}$$

also known as the yield of the so-called "hashing" protocol [16], is negative for $F < 0.81$. To incorporate rate into our utility, we multiply the right-hand side of (17) by R, and for a given route r, define a utility function as follows:

$$U_r^D(R_r, \boldsymbol{w}) = \log\left(R_r D_H(F_r^{\text{e2e}})\right), \tag{18}$$

where F_r^{e2e} is the end-to-end fidelity of entangled states produced for route r. Above, composition by the log is motivated by numerous examples encountered in classical NUM wherein such an approach enforces a sense of fairness to all simultaneously active communication sessions found in a network. In particular, since $\lim_{x \to 0} \log(x) = -\infty$, we can be sure that each route receives a fraction of the aggregate

[2]*I.e.*, states of the form $F|\Psi^+\rangle\langle\Psi^+| + (1-F)/(d^2-1)(\mathbb{I} - |\Psi^+\rangle\langle\Psi^+|)$ with d specifying state dimension.

distillable entanglement from every link that comprises it. A consequence of encasing D_H by the $\log$ is that we implicitly enforce a sufficiently high end-to-end fidelity, *i.e.*, that of at least ≈ 0.81, thereby circumventing the restriction that the bound (16) holds only for $F \in [1/2, 1]$, while also keeping the formulation simple by not having to set restrictions on the end-to-end fidelity for each route. The utility function (18) is very similar to the approach taken by [28], where the authors explored maximization of quantum information flow over repeater chains in a network; notable differences with our approach are the restriction of the objective's domain to ensure non-zero distillable entanglement, and the use of the $\log$ to introduce a notion of fairness.

B. QNUM based on Secret Key Fraction

Another practical measure of entanglement usefulness is the secret key fraction, which describes the amount of secret key that can be generated from an execution of a QKD protocol (*i.e.*, from a successfully-generated entangled state). We will use the secret key fraction of the BB84 protocol [1], which has a convenient closed form for Werner states:

$$S_{BB84}(w) = \max\left(1 - 2h\left(\frac{1-w}{2}\right), 0\right), \quad (19)$$

where $h(\cdot)$ is the binary entropy function [18], [19], [29]. We define a utility function based on (19) as follows:

$$U_r^S(R_r, \boldsymbol{w}) = \log\left(R_r\left(1 - 2h\left(\frac{1 - w_r^{\text{e2e}}}{2}\right)\right)\right), \quad (20)$$

where w_r^{e2e} denotes the Werner parameter of an end-to-end entangled state on route r.

C. QNUM based on Negativity

The final entanglement measure which we include in our study is that of negativity [20], which for a bipartite mixed state ρ on system AB is defined as

$$\mathcal{N}(\rho) \equiv \frac{||\rho^{T_A}||_1 - 1}{2}, \quad (21)$$

where ρ^{T_A} is the partial transpose of ρ with respect to subsystem A and $||\cdot||_1$ is the trace norm. $\mathcal{N}(\rho)$ is also equal to the absolute value of the sum of the negative eigenvalues of ρ^{T_A}, which reduces to zero for unentangled states. For a Werner state with fidelity F, it can be shown that the negativity corresponds to $F - 1/2$, for $F \in [1/2, 1]$. As mentioned previously, this quantity does not map cleanly to an operational interpretation; nevertheless, it is one of the easiest entanglement measures to compute. We next define a utility function based on (21) as follows:

$$\tilde{U}_r^N(R_r, \boldsymbol{w}) = \log(R_r(F_r^{\text{e2e}} - 1/2)), \quad (22)$$

or equivalently,

$$U_r^N(R_r, \boldsymbol{w}) = \log\left(R_r\left(3\prod_{l \in r} w_l - 1\right)\right), \quad (23)$$
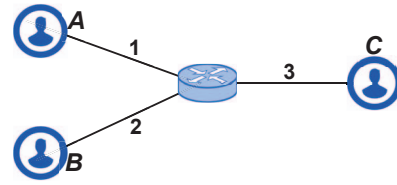

Fig. 2: A simple four-node, three-link topology; node-pairs (A, C) and (B, C) wish to communicate. For all links, $T = 10^{-3}$ s. Links 1 and 2 are 2 km long, while link 3's length is varied in our numerical studies.

which follows from (8), (9), and the fact that we may drop the factor of $1/4$ inside the $\log$ as it has no effect on the optima.

In Appendix A, we show that this utility function is concave – a feature of great mathematical convenience, as we may employ convex optimization techniques with this formulation to reliably find optima. In contrast, the distillable entanglement- and secret key-based utility functions are not in general concave – see Appendix B for proofs.

In all three utility functions, encasing both the rates as well as the entanglement measures within the logarithm is a deliberate design choice, not only for ensuring fairness among the routes, but also for implicitly constraining the domains of these functions so as to ensure non-zero distillable entanglement, secret key rate, and negativity. A consequence, however, is that utility functions (18) and (20) have stringent requirements on end-to-end fidelity. As mentioned previously, D_H is negative for fidelities below 0.81; this implies that any solution to the utility maximization would result in *every* route receiving entanglement of at least this quality, quite likely resulting in low rates R_r across the network. A similar property applies to U^S: to see this, let us examine the function $1 - 2h(p)$, where the relevant interval on p is $[0, 1/2]$ since $w \in [0, 1]$. This quantity is negative for p less than approximately 0.11, which translates to a fidelity of 0.84 – an even stricter requirement than we had with distillable entanglement. In contrast, the negativity-based utility function merely requires end-to-end fidelities of at least $1/2$. What one would expect, then, is for U^D and U^S to produce higher-quality entanglement at low rates, while the opposite holds for U^N. These properties are useful to keep in mind when choosing a utility: the former two, for instance, may be more suitable for QKD or applications with high end-to-end fidelity requirements, while the latter may be more suitable to applications that are sensitive to entanglement inter-arrival times (jitter), such as BQC.

V. NUMERICAL INVESTIGATION

We apply the distillable entanglement- (DE), secret key fraction- (SKF), and negativity- (NGTV) based QNUM constructions of the previous section to a number of network topologies chosen so as to expose important behavioral properties of our utility formulations. In particular, we opt for

1242

topologies wherein symmetry may be exploited in the sense that all routes derive the same amount of utility from the network – this allows us to more easily predict and interpret performance metrics of interest. In such highly symmetric network topologies, the optimization is fairly easy to execute via grid search on few parameters (in our examples, only two – *e.g.*, the Werner parameter of a metropolitan-area link and a backbone link). In general, however, optimization may be performed using more sophisticated, algorithmic methods. For the topology in Figure 2, for instance, we use a bound-constrained augmented Lagrangian approach [30] with quadratic penalty functions to enforce the equality constraints (12). For the subproblems, we use the projected gradient descent method with backtracking line search to adaptively choose the step lengths [31].

The first topology that we study is depicted in Figure 2 and is a three-link, three-user network with a single repeater node set up so that the two users on the left wish to communicate with the user on the right. Links 1 and 2 (connecting the users on the left to the repeater) are identical and 2 km long, while link 3's length is varied from 2 to 180 km. All links have constant repetition times of $T = 10^{-3}$ s and the non-fiber inefficiencies coefficient c is set to one. There are thus two "routes" (communication sessions) in the network, resulting in the following optimization problem:

$$\max\ U_1(R_1, \boldsymbol{w}) + U_2(R_2, \boldsymbol{w}) \text{ subject to}$$
$$R_i = d_i(1 - w_i), i = 1, 2, \qquad (24)$$
$$R_1 + R_2 = d_3(1 - w_3),$$

where the utilities make use of post-swap Werner parameters of route 1/2, which are given by $w_1 w_3 / w_2 w_3$, respectively. Figure 3 (left) presents individual elementary link-level fidelities for this topology. As one would expect, since link 3 must service two communication sessions, while links 1 and 2 service only one each, the former's fidelity is in general lower in order to support the additional rate demand. Links 1 and 2, on the other hand, have higher fidelities to compensate for link 3's lower one. Moreover, as link 3's length increases, links 1 and 2 are further coerced to produce higher-quality entanglement to make up for the loss in fidelity of link 3's entanglement.

The resulting aggregate network utility (Figure 3 center) decreases approximately linearly as a function of link 3's length, evidently due to the network's lowered capability of producing entanglement of high-quality, or at high rates. The rightmost panel of the figure presents end-to-end fidelities and rates as a function of link 3's length, for each individual route (as previously mentioned, all routes receive the same QoS due to the symmetry of the problem). We observe here that all three utilities appear to maintain a fairly steady end-to-end entanglement fidelity, while sacrificing route rates, which appear to decrease at an exponential rate with the link length.

A striking observation so far is that DE- and SKF-based utility formulations appear to be in close agreement – they produce similar solutions and subsequently have comparable

notions of utility. The negativity-based formulation, on the other hand, puts rate and fidelity on a much more equal footing, thereby allowing considerably lower end-to-end fidelities. Of the three utilities, the SKF-based one is most stringent on fidelity (albeit only by a small amount relative to the DE-based utility).

A generalization of the topology in Figure 2 is the "clients-server" topology illustrated in Figure 4. Here, multiple users connected to a repeater node form a rudimentary metropolitan-area network. All user-repeater links are identical, have 15 km lengths, and repetition times $T = 10^{-3}$ s. The repeater-server link, which has the potential to become a bottleneck, is a longer-distance link of 100 km and has $T = 10^{-4}$ s. For all links, $c = 0.1$. The QNUM formulation is similar to that of (24), with the exception that we sum over all n route utilities within the objective, and the repeater-server link (denoted by the subscript Bb) yields a constraint of the form

$$\sum_{i=1}^{n} R_i = d_{\text{Bb}}(1 - w_{\text{Bb}}).$$

Figure 5 (left) presents the optimal link-level fidelities as a function of the number of users in the metro-area network, with "M-Link" denoting metropolitan-area links and "Bb-Link" denoting the repeater-server link. When there is only one user in the metro-area network, the repeater-server link is relatively unburdened, and due to $d_{\text{Bb}} > d_{\text{metro}}$, where d_{metro} is any metropolitan-area link's parameter, produces higher-fidelity entanglement than metro-area links. As the number of nodes grows, increasing the load on "Bb-link", we observe a crossover – the link's fidelity must decrease to accommodate the additional demand, causing metro-area links to compensate in fidelity.

Figure 5 (center) presents the resulting network utility, where we observe a new behavior: up to a threshold, the aggregate utility increases with the number of nodes. A possible explanation for this is that for a sufficiently small number of nodes, the network may be underutilized. Consider a scenario where a powerful backbone link, for example, is able to generate high-fidelity entanglement at a relatively high rate. An individual route that utilizes such a link may not be able to keep up with this rate of entanglement generation without excessively impairing the fidelity of the states it generates. If the number of routes utilizing the backbone link increases (up to a threshold which depends on the utility function and other network parameters), the aggregate utility may increase even if individual route utilities are lower than the utility of the single-route scenario. Beyond the threshold, the backbone link becomes overburdened and begins to sacrifice fidelity, resulting in lower aggregate utilities.

Figure 5 (right) presents individual route end-to-end fidelities and rates – as with the topology example in Figure 2, we observe rapid decreases in the rates when resources become scarcer in the network (*i.e.*, as the number of users increases in the present example). As the repeater-server link begins to service more users, thereby splitting its aggregate entanglement generation rate among a larger number of routes, we
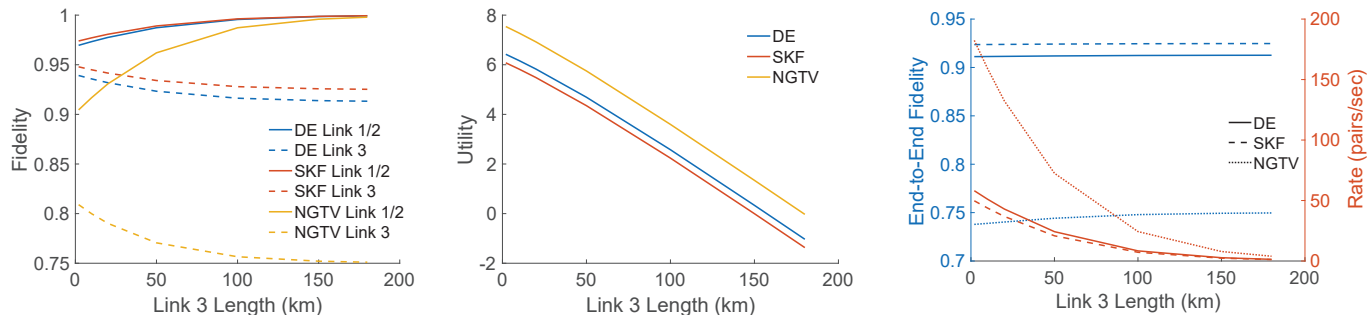
Fig. 3: Optimal link-level fidelity values, resulting utilities, end-to-end fidelities, and rates for the three-link network in Figure 2, as a function of the longest link's (link 3) length.
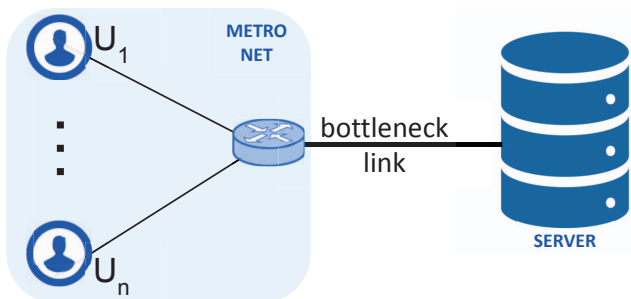

Fig. 4: A clients-server topology with n users and one repeater node connecting them to the server via a bottleneck link. Links that connect users to the network are identical and 15 km with $T = 10^{-3}$ s, while the bottleneck link is 100 km with $T = 10^{-4}$ s. The number of users is varied in our numerical studies.

also observe an increase in end-to-end fidelity – a consequence of metro-area links being coerced to generate entanglement at lower rates.

We next study a simple line network is shown in Figure 6. Here, links 1 and 3 are both 15 km and have $T = 10^{-3}$ s, while link 2 has $T = 10^{-4}$ s and its length is varied between 5 and 300 km. For all links, $c = 0.1$. Figure 7 (left) presents the optimal link-level fidelity as a function of link 2's length, where the crossover effect is even more apparent than that of the previous example. In the center panel of the figure, we observe a monotonic decrease in the network utility – a natural consequence of link 2's decreased ability of producing high-fidelity states. In the right panel of the figure, we again see little deviation in an individual route's end-to-end fidelity, while rates now exhibit a more interesting behavior – both concave as well as convex profiles in the rate curves.

Our final numerical study is that of the dumbbell topology as illustrated in Figure 8. Here, each metropolitan-area network has $n/2$ users, each of which has a communication session with a user in the other network. A backbone link connects the two metropolitan areas. All metro-area links are identical

15 km links with $T = 10^{-3}$ s, while the backbone is a 100 km link with $T = 10^{-4}$ s. $c = 0.1$ for all links. Figure 9 (left) presents the optimal metro-area ("M-Link") and backbone ("Bb-Link") fidelities, as the number of nodes in each metro-area varies. We again see the crossover effect, as well as the increase in aggregate utility (up to a threshold) similar to that of the clients-server network example (center panel). Finally, the right panel of the figure presents a familiar behavior: fairly constant end-to-end fidelities, with rates paying the price.

VI. CONCLUSION

The aim of this work has been to explore ways of adapting classical NUM to quantum networked settings. To this end, we proposed three candidate quantum network utility functions, each based on a different entanglement measure. We applied QNUM to a number of network topologies to solve the resource allocation problem for networks that have inherent rate-fidelity tradeoffs during the elementary link entanglement generation process. We have found that formulations incorporating distillable entanglement and secret key fraction are quite similar in their behavior, and that they both place a high value on end-to-end fidelity of entanglement. Our third formulation, based on entanglement negativity, allows for lower end-to-end fidelities, and as a result yields higher rates to user-pairs wishing to share entanglement. These contrasting behaviors raise questions about suitable QNUM constructions tailored to specific distributed quantum applications.

APPENDIX A
CONCAVITY OF THE NEGATIVITY-BASED UTILITY FUNCTION

In this appendix, our goal is to show that the utility function $U_r^N(R_r, \boldsymbol{w}) = \log\left(R_r \left(3 \prod_{l \in r} w_l - 1 \right) \right)$ defined on the domain

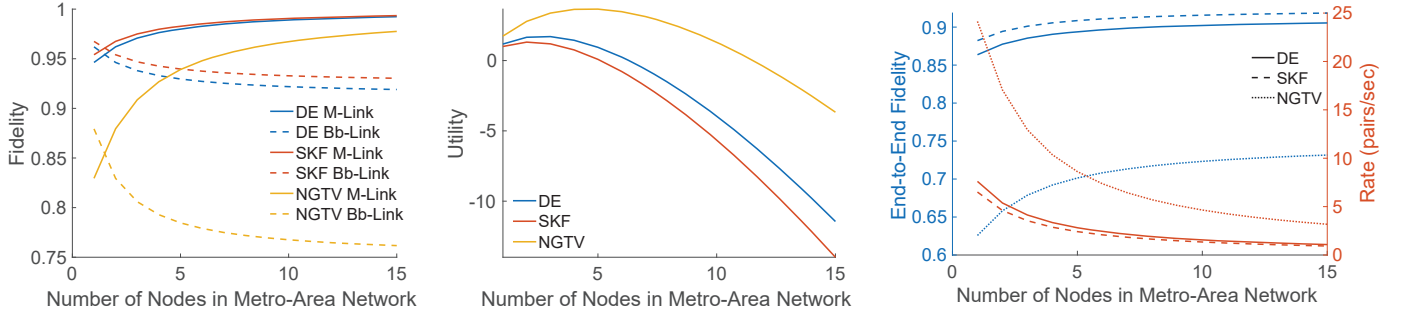$$\left\{ R_r > 0, \boldsymbol{w} > \boldsymbol{0} : \prod_{l \in r} w_l > \frac{1}{3} \right\} \qquad (25)$$

Fig. 5: Optimal link-level fidelity values, resulting utilities, end-to-end fidelities, and rates for the client-server network in Figure 4, as a function of the number of users.
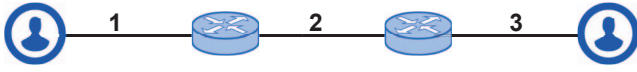

Fig. 6: A line topology with two users and two repeater nodes. Links 1 and 3 are identical, while link 2's length is varied for our numerical studies.

is concave. To do so, it suffices to show that the function $f(\boldsymbol{x}) = -\log(\prod_i x_i - 1)$, where $x_i > 0$ are such that $\prod x_i > 1$, is convex. The reason for this is as follows:

$$\log\left(R_r\left(3\prod_{l\in r} w_l - 1\right)\right) = \log R_r + \log\left(\prod_{l\in r} \sqrt[|r|]{3}w_l - 1\right),$$
(26)

where $|r|$ is the number of links in route r. It is known that a sum of concave functions is also concave. The argument then follows by letting $x_l \equiv \sqrt[|r|]{3}w_l$, where we know that $x_l > 0$ since w_l are strictly positive, and we also know that $\prod_l x_l > 1$ since by (25), $3\prod_l w_l > 1$.

Our first goal is to prove that the domain of this function, *i.e.*, the set $S \equiv \{\boldsymbol{x} > \boldsymbol{0} : \prod x_i > 1\}$, is a convex set. For this we must ensure that given any two points $\boldsymbol{x} \neq \boldsymbol{y} \in S$, $\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}$ is also in S, for any $\theta \in [0,1]$. *I.e.*, we require

$$\prod_{i=1}^{n}(\theta x_i + (1-\theta)y_i) > 1,$$
(27)

where n is the number of elements in $\boldsymbol{x}$ (or $\boldsymbol{y}$). Expanding the left-hand side of (27), we have

$$\theta^n \prod_{i=1}^{n} x_i + (1-\theta)^n \prod_{i=1}^{n} y_i +$$
$$\sum_{k=1}^{n-1} \theta^k (1-\theta)^{n-k} \sum_{i_1=1}^{n-k+1} \sum_{i_2=i_1+1}^{n-k} \cdots \sum_{i_k=i_{k-1}+1}^{n} x_{i_1} x_{i_2} x_{i_k} \prod_{j\neq i_1,\ldots,i_k} y_j$$
$$> \theta^n \prod_{i=1}^{n} x_i + (1-\theta)^n \prod_{i=1}^{n} y_i +$$

$$\sum_{k=1}^{n-1} \theta^k (1-\theta)^{n-k} \binom{n}{k} \left(\prod_{i=1}^{n} x_i^{\binom{n-1}{k-1}} \prod_{j=1}^{n} y_j^{\binom{n-1}{k}}\right)^{1/\binom{n}{k}},$$
(28)

where (28) follows from the AM-GM inequality and the fact that for a given coefficient $\theta^k(1-\theta)^{n-k}$ and variable x_i, there are $\binom{n-1}{k-1}$ cross terms that involve x_i, and for a given variable y_i, there are $\binom{n-1}{n-k-1} = \binom{n-1}{k}$ cross terms that involve it. Next, using our knowledge of $\boldsymbol{x}$ and $\boldsymbol{y}$ both belonging to S, we obtain

$$\prod_{i=1}^{n}(\theta x_i + (1-\theta y_i)) > \theta^n + (1-\theta)^n + \sum_{k=1}^{n-1}\binom{n}{k}\theta^k(1-\theta)^{n-k}$$
$$= (\theta + (1-\theta))^n = 1,$$

proving (27).

To complete the proof that $f(\boldsymbol{x})$ is convex, we show that its Hessian is positive definite. The first- and second-order partials of f are as follows:

$$\frac{\partial f}{\partial x_j} = -\frac{\prod_{i\neq j} x_i}{\prod_i x_i - 1},$$
(29)

$$\frac{\partial^2 f}{\partial x_j^2} = \left(\frac{\prod_{i\neq j} x_i}{\prod_i x_i - 1}\right)^2, \quad \text{and for } j \neq k,$$
(30)

$$\frac{\partial^2 f}{\partial x_j \partial x_k} = \frac{-\prod_{i\neq j,k} x_i \left(\prod_i x_i - 1\right) + \prod_{i\neq j} x_i \prod_{i\neq k} x_i}{\left(\prod_i x_i - 1\right)^2}$$
(31)

$$= \frac{-\prod_{i\neq j,k} x_i \prod_i x_i + \prod_{i\neq j,k} x_i + \prod_{i\neq j} x_i \prod_{i\neq k} x_i}{\left(\prod_i x_i - 1\right)^2}$$
(32)

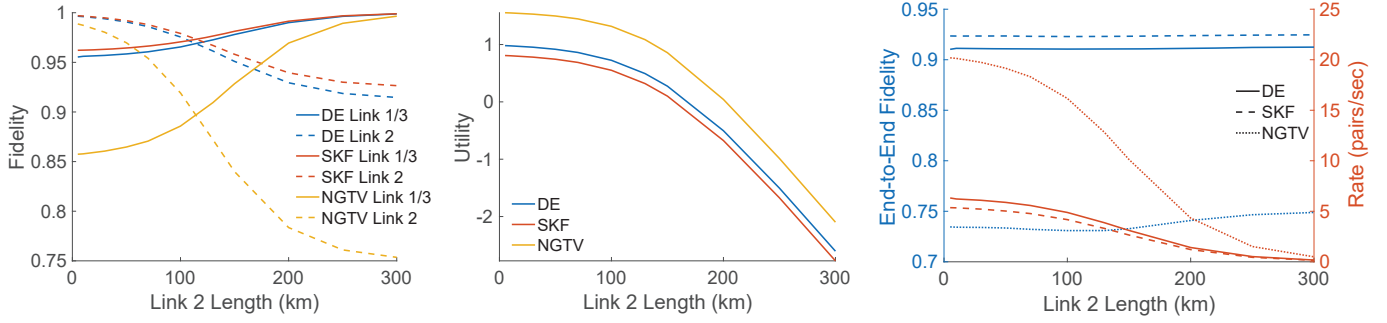$$= \frac{\prod_{i\neq j,k} x_i}{\left(\prod_i x_i - 1\right)^2}.$$
(33)

Fig. 7: Optimal link-level fidelity values, resulting utilities, end-to-end fidelities, and rates for the line network in Figure 6, as a function of the middle link's (link 2) length.
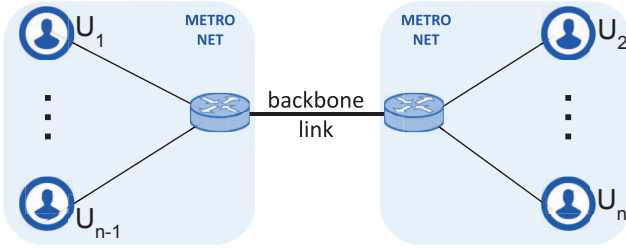

Fig. 8: A dumbbell topology with n users and two repeater nodes connected by a backbone link. Links that connect users to the network are identical. User pairs $(U_1, U_2), \ldots, (U_{n-1}, U_n)$ wish to communicate. The number of communicating user pairs is varied in our numerical studies.

To simplify notation, we introduce

$$A = \prod_i x_i - 1, \quad B_j = \prod_{i \neq j} x_i, \quad \text{and} \quad B_{j,k} = \prod_{i \neq j,k} x_i,$$

so that the second-order partials are $\partial^2 f / \partial x_j^2 = B_j^2 / A^2$ and $\partial^2 f / \partial x_j \partial x_k = B_{j,k} / A^2$. If x has dimension n, the Hessian of $f(x)$ is given by

$$H = \frac{1}{A^2} \begin{bmatrix} B_1^2 & B_{1,2} & \cdots & B_{1,n} \\ B_{1,2} & B_2^2 & \cdots & B_{2,n} \\ \vdots & & \ddots & \vdots \\ B_{1,n} & \cdots & B_{n-1,n} & B_n^2 \end{bmatrix}. \quad (34)$$

For positive definiteness of H, we require that for any non-zero vector y of dimension n, $y^T H y > 0$. I.e., it must be that

$$\frac{y^T}{A^2} \begin{bmatrix} y_1 B_1^2 + \sum_{i \neq 1} y_i B_{1,i} \\ y_2 B_2^2 + \sum_{i \neq 2} y_i B_{2,i} \\ \vdots \\ y_n B_n^2 + \sum_{i \neq n} y_i B_{n,i} \end{bmatrix} > 0.$$

Since A and therefore A^2 are strictly positive, we end up with

$$y_1^2 B_1^2 + y_1 \sum_{i \neq 1} y_i B_{1,i} + y_2^2 B_2^2 + y_2 \sum_{i \neq 2} y_i B_{2,i} + \cdots \quad (35)$$
$$+ y_n^2 B_n^2 + y_n \sum_{i \neq n} y_i B_{n,i} > 0,$$

$$\sum_{i=1}^n \left(y_i^2 B_i^2 + y_i \sum_{j \neq i} y_j B_{i,j} \right) > 0. \quad (36)$$

Here, it is useful to introduce $P \equiv \prod_i x_i$, and note that $B_i = P/x_i$ and $B_{i,j} = P/(x_i x_j)$. Thus, (36) becomes

$$\sum_{i=1}^n \left(y_i^2 \frac{P^2}{x_i^2} + \sum_{j \neq i} y_i y_j \frac{P}{x_i x_j} \right) > 0. \quad (37)$$

Letting $\gamma_i = y_i / x_i$ and dividing both sides of (37) by the strictly positive P, we obtain

$$\sum_{i=1}^n \left(P \gamma_i^2 + \sum_{j \neq i} \gamma_i \gamma_j \right) > 0. \quad (38)$$

Recalling that $P > 1$, we note that $P \gamma_i^2 > \gamma_i^2$, $\forall i$. Thus, it would suffice to show that

$$\sum_{i=1}^n \left(\gamma_i^2 + \sum_{j \neq i} \gamma_i \gamma_j \right) > 0, \quad (39)$$

$$\sum_{i=1}^n \sum_{j=1}^n \gamma_i \gamma_j > 0, \quad (40)$$

$$\left(\sum_{i=1}^n \gamma_i \right)^2 > 0, \quad (41)$$

which clearly holds. $\qquad\square$

APPENDIX B
NON-CONCAVITY OF DISTILLABLE ENTANGLEMENT- AND SECRET KEY-BASED UTILITY FUNCTIONS

To show that utility functions (18) and (20) are not in general concave, it suffices to prove this for the simplest
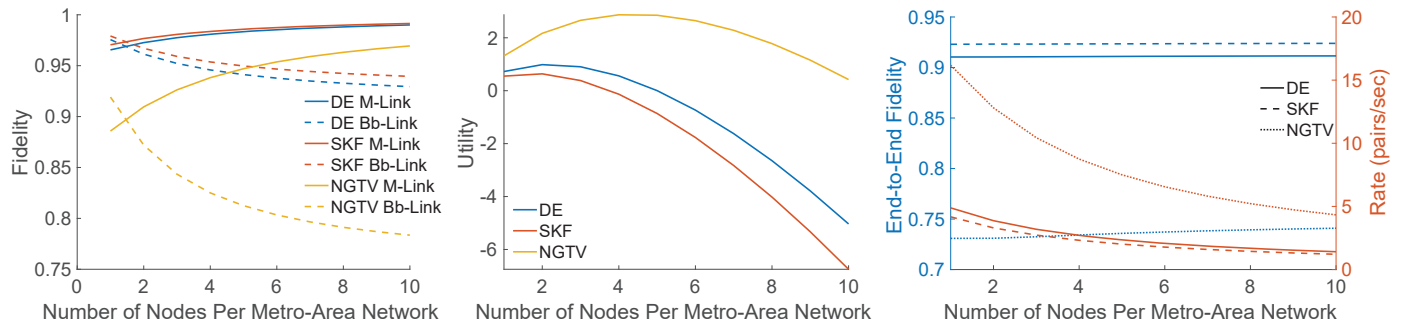
Fig. 9: Optimal link-level fidelity values, resulting utilities, end-to-end fidelities, and rates for the dumbbell network in Figure 8, as a function of the number of nodes in each metropolitan-area network.

example of a single link, resulting in utilities of two variables: a rate R and a Werner parameter w. We then must determine whether the functions $f(R, w) := -U^D(R, w)$ and $g(R, w) := -U^S(R, w)$ are convex. The second-order partials of f and g with respect to w, are

$$\frac{\partial^2 f}{\partial w^2} = \left(\frac{3}{4 D_H(F)} \log_2 \left(\frac{3w + 1}{1 - w} \right) \right)^2 - \frac{3}{D_H(F) \log(2)(3w + 1)(1 - w)}, \quad (42)$$

$$\frac{\partial^2 g}{\partial w^2} = \left(\frac{1}{S_{\text{BB84}}(w)} \log_2 \left(\frac{1 + w}{1 - w} \right) \right)^2 - \frac{2}{S_{\text{BB84}}(w) \log(2)(1 + w)(1 - w)}, \quad (43)$$

where F is given in (8), D_H is defined in (17), and S_{BB84} is defined in (19). These partials evaluate to negative values for values of $w > 0.97$ (equivalently, fidelities above 0.98 – well within the acceptable domain of U^D and U^S): for instance, (42) yields -3.19 and (43) yields -2.34 when $w = 0.97$. Since these partials are the leading principal minors of the Hessians of f and g, we conclude that these functions are not convex. □

ACKNOWLEDGMENT

REFERENCES

[1] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," in *Proc. of IEEE Int. Conf. on Comp. Sys. and Signal Proc.*, 1984.

[2] A. K. Ekert, "Quantum cryptography and bell's theorem," in *Quantum Measurements in Optics.* Springer, 1992, pp. 413–418.

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

[4] R. Jozsa, D. S. Abrams, J. P. Dowling, and C. P. Williams, "Quantum clock synchronization based on shared prior entanglement," *Physical Review Letters*, vol. 85, no. 9, p. 2010, 2000.

[5] 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.

[6] J. F. Fitzsimons and E. Kashefi, "Unconditionally verifiable blind quantum computation," *Physical Review A*, vol. 96, no. 1, p. 012303, 2017.

[7] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, 2018.

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

[9] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, 1998.

[10] C. Jin, D. X. Wei, and S. H. Low, "FAST TCP: motivation, architecture, algorithms, performance," in *IEEE INFOCOM 2004*, vol. 4. IEEE, 2004, pp. 2490–2501.

[11] M. Zukowski, A. Zeilinger, M. A. Horne, and A. K. Ekert, "" Event-ready-detectors" Bell experiment via entanglement swapping." *Physical Review Letters*, vol. 71, no. 26, 1993.

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

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

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

[15] P. C. Humphreys, N. Kalb, J. P. Morits, R. N. Schouten, R. F. Vermeulen, D. J. Twitchen, M. Markham, and R. Hanson, "Deterministic delivery of remote entanglement on a quantum network," *Nature*, vol. 558, no. 7709, pp. 268–273, 2018.

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

[17] E. M. Rains, "A semidefinite program for distillable entanglement," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2921–2933, 2001.

[18] P. W. Shor and J. Preskill, "Simple proof of security of the BB84 quantum key distribution protocol," *Physical review letters*, vol. 85, no. 2, p. 441, 2000.

[19] H.-K. Lo, H. F. Chau, and M. Ardehali, "Efficient quantum key distribution scheme and a proof of its unconditional security," *Journal of Cryptology*, vol. 18, no. 2, pp. 133–165, 2005.

[20] G. Vidal and R. F. Werner, "Computable measure of entanglement," *Physical Review A*, vol. 65, no. 3, p. 032314, 2002.

[21] R. Srikant and L. Ying, *Communication networks: an optimization, control, and stochastic networks perspective.* Cambridge University Press, 2013.

[22] L. Jiang and J. Walrand, "A distributed csma algorithm for throughput and utility maximization in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 960–972, 2009.

[23] F. Shams, G. Bacci, and M. Luise, "A survey on resource allocation techniques in OFDM (A) networks," *Computer Networks*, vol. 65, pp. 129–150, 2014.

[24] L. S. Brakmo, S. W. O'Malley, and L. L. Peterson, "TCP Vegas: New techniques for congestion detection and avoidance," in *Proceedings of the conference on Communications architectures, protocols and applications*, 1994, pp. 24–35.

[25] Y. Yi and M. Chiang, "Stochastic network utility maximisation—a tribute to Kelly's paper published in this journal a decade ago," *European Transactions on Telecommunications*, vol. 19, no. 4, pp. 421–442, 2008.

[26] F. Rozpedek, "Building blocks of quantum repeater networks," Ph.D. dissertation, 2019.

[27] Á. G. Iñesta, G. Vardoyan, L. Scavuzzo, and S. Wehner, "Optimal entanglement distribution policies in homogeneous repeater chains with cutoffs," *arXiv preprint arXiv:2207.06533*, 2022.

[28] M. Victora, S. Krastanov, A. S. de la Cerda, S. Willis, and P. Narang, "Purification and entanglement routing on quantum networks," *arXiv preprint arXiv:2011.11644*, 2020.

[29] B. Li, T. Coopmans, and D. Elkouss, "Efficient optimization of cut-offs in quantum repeater chains," in *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, 2020, pp. 158–168.

[30] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[31] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.