The Capacity of Classical Summation over a Quantum MAC with Arbitrarily Distributed Inputs and Entanglements

Yuhang Yao, Syed A. Jafar
Center for Pervasive Communications and Computing (CPCC)
University of California Irvine, Irvine, CA 92697
Email: {yuhangy5, syed}@uci.edu *

Abstract

The Σ-QMAC problem is introduced, involving S servers, K classical ($\mathbb{F}_d$) data streams, and T independent quantum systems. Data stream $\mathsf{W}_k, k \in [K]$ is replicated at a subset of servers $\mathcal{W}(k) \subset [S]$, and quantum system $\mathcal{Q}_t, t \in [T]$ is distributed among a subset of servers $\mathcal{E}(t) \subset [S]$ such that Server $s \in \mathcal{E}(t)$ receives subsystem $\mathcal{Q}_{t,s}$ of $\mathcal{Q}_t = (\mathcal{Q}_{t,s})_{s\in\mathcal{E}(t)}$. Servers manipulate their quantum subsystems according to their data and send the subsystems to a receiver. The total download cost is $\sum_{t\in[T]}\sum_{s\in\mathcal{E}(t)} \log_d |\mathcal{Q}_{t,s}|$ qudits, where $|\mathcal{Q}|$ is the dimension of $\mathcal{Q}$. The states and measurements of $(\mathcal{Q}_t)_{t\in[T]}$ are required to be separable across $t \in [T]$ throughout, but for each $t \in [T]$, the *subsystems* of $\mathcal{Q}_t$ can be prepared initially in an arbitrary (independent of data) entangled state, manipulated arbitrarily by the respective servers, and measured jointly by the receiver. From the measurements, the receiver must recover the sum of all data streams. Rate is defined as the number of dits ($\mathbb{F}_d$ symbols) of the desired sum computed per qudit of download. The capacity of Σ-QMAC, i.e., the supremum of achievable rates is characterized for arbitrary data replication and entanglement distribution maps $\mathcal{W}, \mathcal{E}$. For example, in the symmetric setting with $K = \binom{S}{\alpha}$ data-streams, each replicated among a distinct α-subset of $[S]$, and $T = \binom{S}{\beta}$ quantum systems, each distributed among a distinct β-subset of $[S]$, the capacity of the Σ-QMAC is $\frac{1}{\beta T}\sum_{\gamma=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)} \min(\beta, 2\gamma) \cdot \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma}$. Coding based on the N-sum box abstraction is optimal in every case. Notably, for every $S \neq 3$ there exists an instance of the Σ-QMAC where S-party entanglement is necessary to achieve the fully entangled capacity.

*To appear in part at IEEE GLOBECOM 2023 [1].

1 Introduction

Entanglement is arguably the most counter-intuitive aspect of quantum systems. Quantum entanglement enables correlations that are classically impossible. These correlations can be exploited for improvements in the efficiency of quantum communication and computation networks. Understanding the fundamental limits of quantum entanglement phenomena is therefore essential to gauge the potential of the much-anticipated quantum internet of the future [2–4]. However, even quantifying the amount of entanglement is highly non-trivial, especially when the entanglement is distributed among *many* parties. Unlike bipartite entanglement which is relatively well understood — a fundamental understanding of multi-party[1] entanglement remains elusive. Numerous fundamentally distinct measures have been explored thus far [5], including the Schmidt measure [6, 7], the trace-squared or the entropy of the reduced density matrix [8], the tangle [9], the entanglement of formation [10], majorization-based entanglement monotones [11, 12], geometric measures [13–15], and specialized notions such as absolute maximal or genuine multiparty entanglement [16, 17].

Given the lack of a universal measure, an alternative is to quantify multi-party entanglement indirectly in terms of its utility as a resource [18], e.g., by the gains in communication efficiency[2] that are made possible by quantum entanglement for accomplishing various classical multiparty computation tasks. This leads to a broad array of scenarios [20]. To list a few, the communication may be interactive or it may follow a non-interactive model such as *simultaneous message passing* (SMP) [20–23], inputs may be distributional or prior-free, the output may be a total function, a partial function or a relation, the computation may be exact or probabilistic, limited to a single-instance or with parallel-repetitions and batch processing, and with/without security/privacy constraints and/or classical common randomness [24, 25]. Much effort has traditionally been aimed at finding tasks that gain a lot from quantum entanglement in terms of communication complexity measures [20, 26]. A limitation of this approach is that the tasks thus identified may turn out to be artificial. For example, the celebrated Deutsch-Jozsa algorithm [27] translates into a computation task in a 2 user quantum SMP setting [22, 23] where entanglement shows an exponential advantage. However, such a computation task is seldom encountered in practice. A complementary approach that we explore in this work, is to focus instead on some elementary computation tasks that are quite natural, such as linear computations, and explore the gains in efficiency due to quantum entanglements for such tasks. Specifically, we study an elemental setting, called the Σ-QMAC, to be described shortly, where the computation task is simply a finite-field summation task, over an ideal (noise-free) quantum multiple access network with arbitrarily distributed inputs and entanglements. Since the gains may be modest, a finer accounting of efficiency, such as the exact information theoretic capacity, is needed.

The pursuit of exact capacity faces numerous challenges – 1) sharp capacity characterizations are quite rare even for classical communication networks when many parties are involved, 2) computation networks tend to be even less tractable than communication networks for information theoretic analysis, and 3) the quantum setting further compounds the difficulty of any such endeavor. Indeed the foremost challenge in pursuing this direction is to identify formulations that are both insightful from a multiparty quantum entanglement perspective and also information theoretically tractable. Our choice of the Σ-QMAC setting draws inspiration from the following

[1] We refer to entanglement among N-parties as bipartite if $N = 2$, and multiparty if $N > 2$.

[2] Indeed, it is conjectured that communication efficiency may provide a concise information-theoretic axiomatic basis for characterizing quantum mechanics [19].

observations — 1) the many-to-one (multiple access (MAC)) setting is among the most tractable in network information theory, 2) the capacity of finite field linear computations in noiseless settings, while still open in general, has seen much progress in the network coding literature, in particular elegant solutions have been found for the case of scalar linear computations (i.e., sum computations) [28–32], and 3) the stabilizer formalism [33–35] from quantum error correction lends itself nicely to linear black-box abstractions for the quantum multiple access channel (QMAC) [36], and has been instrumental to recent advances in quantum private information retrieval (QPIR) [36–41] that implicitly involve optimal specialized linear computations. The Σ-QMAC setting is described next.

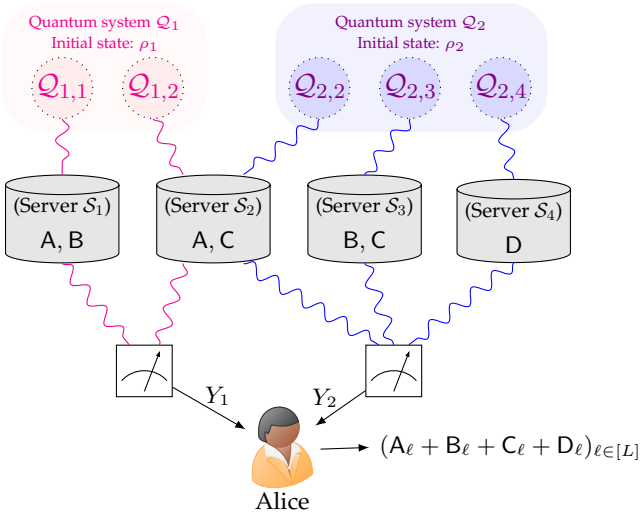

Figure 1: A Σ-QMAC setting, with $K = 4$ data streams $(\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D})$, $S = 4$ servers $(\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4)$, and $T = 2$ quantum systems $(\mathcal{Q}_1, \mathcal{Q}_2)$. The data replication map $\mathcal{W} = (\{1,2\}, \{1,3\}, \{2,3\}, \{4\})$ specifies that data stream A is replicated at Servers $\mathcal{S}_1, \mathcal{S}_2$; B at $\mathcal{S}_1, \mathcal{S}_3$; C at $\mathcal{S}_2, \mathcal{S}_3$; and data stream D is available only at $\mathcal{S}_4$. The entanglement distribution map $\mathcal{E} = (\{1,2\}, \{2,3,4\})$ is such that entangled subsystems of $\mathcal{Q}_1$ are distributed to Servers $\mathcal{S}_1, \mathcal{S}_2$, and entangled subsystems of $\mathcal{Q}_2$ are distributed to Servers $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$.

In a Σ-QMAC setting (see Section 4 for a formal definition), a user, say Alice, wants to compute the *sum* of K classical data streams (comprised of d-ary symbols (*dits*) from a finite field $\mathbb{F}_d$) that are replicated across various subsets of S servers (cf. graph-based replication [42–44]) according to an arbitrary data replication map $(\mathcal{W})$. Independent of the data streams, T quantum systems are prepared, and subsystems of each quantum system are distributed to various subsets of servers (called cliques) according to an arbitrary entanglement distribution map $(\mathcal{E})$. The states and the eventual measurements of different quantum systems must remain separable throughout, but the *subsystems* of each quantum system are in general entangled even as they are distributed to different servers within that clique, thus allowing such a clique of servers to exploit their quantum entanglement. The servers locally encode their classical data into their quantum subsystems, maintaining separation among different quantum systems associated with different cliques, and send them to Alice, who does separate measurements on each quantum system. From the measurement outcome, Alice must be able to recover the desired sum. The computation rate is the

number of dits of the desired sum computed by Alice per qudit of download.[3] An example is illustrated in Fig. 1. If Alice is able to compute L *dits* of the desired sum ($\mathsf{A}+\mathsf{B}+\mathsf{C}+\mathsf{D}$ in Fig. 1) with total communication cost N *qudits* then the rate achieved is L/N (dits/qudit). The capacity C is the supremum of achievable rates.

In order to quantify the utility of multiparty quantum entanglements, the key figure of merit in the Σ-QMAC is the multiplicative gain in capacity that is enabled by entanglement, relative to the corresponding unentangled setting. In the literature such a gain is known as *distributed superdense coding gain* [18, 36–40, 45–49]. Quantifying the utility of multiparty entanglements by characterizing the distributed superdense coding (DSC) gain in the Σ-QMAC is the immediate focus of this work. The broader motivation is that success in the Σ-QMAC setting may pave the way for future studies of *general* linear computation tasks that shed further light on the fundamental limits of the utility of multiparty quantum entanglement.

2 Significance of the Σ-QMAC and relationship to prior works

The underlying quantum multiple access (QMAC) communication model in the Σ-QMAC is similar to what is known in the literature as *simultaneous message passing* (SMP) model with quantum messages [20, 21, 23] . A noteworthy distinction is that the SMP model is typically studied from a communication complexity perspective which does not allow batch processing, whereas since our perspective is information theoretic, batch processing is not only allowed, it is essential to our problem formulation. For brevity, and to underscore the information theoretic perspective, we say QMAC when we mean an SMP model with quantum messages and batch processing.

In the Σ-QMAC the inputs are prior-free, the desired output is exact, the channel is idealized, and the number of servers can be arbitrarily large. The significance of these assumptions is explained as follows. The prior-free model is desirable for computation problems, because unlike conventional communication problems where independent messages can be separately compressed to their entropy limit to yield uniform data, for computation problems the compression of inputs cannot be taken for granted as it changes the nature of the computation. The prior-free model is also quite robust as the results are not limited to one data distribution or another. Exact computation goes hand-in-hand with the assumption of prior-free inputs, because probabilistic bounds on errors are less meaningful when no particular distribution is assumed on the data. The idealized communication channel which transmits qudits noise-free from the servers to Alice is also important in this regard. The idealized channel ensures that the capacity of the Σ-QMAC reflects only the fundamental limitations of the quantum-entanglements for the chosen task, and not other artifacts that arise out of channel-imposed limitations. Furthermore, since we wish to explore multi-partite entanglements among a large number of parties it is important that we explore Σ-QMAC settings with arbitrarily large number of entangled servers. These considerations highlight the essential distinctions that separate our work from other interesting research directions pursued, for example in [50], [51] and [52], that explore ϵ-error sum computation over a QMAC with correlated data streams and noisy quantum channels, albeit with only 2 servers (transmitters).

In the Σ-QMAC we allow arbitrary entanglements across the subsystems of each quantum system, but quite importantly, we do *not* allow entanglements across systems. We require that

[3] In contrast to a *dit*, which is a classical d-ary symbol, a *qudit*, short for a quantum-dit, represents a d-dimensional quantum system. For $d=2$ these are the common 'bit' and 'qubit,' respectively.

strict separation between quantum systems be preserved throughout. The significance of this assumption is that it allows us to determine whether multiparty entanglement is in fact necessary to achieve the DSC gain for a given Σ-QMAC setting. For example, suppose we wish to determine the DSC gain achievable in a Σ-QMAC with only bipartite entanglements. To this end we allow every pair of servers to share unlimited amount of bipartite entanglements, but no multiparty entanglements are initially provided to the servers. Without the separate processing constraint it may still be possible for the Σ-QMAC capacity to benefit from multiparty entanglement, e.g., due to implicit or explicit fusion [53] among bipartite quantum systems through joint processing and measurements. The DSC gain thus achieved could not be categorically attributed to only bipartite entanglement. With the separate processing constraint on the other hand, the resulting DSC gain can *only* be attributed to bipartite entanglements. In fact, it is crucial for our motivation of understanding fundamental limits of multiparty entanglements that we are able to convincingly determine whether the DSC gains necessarily require multiparty entanglement, and furthermore to be able to distinguish between DSC gains possible with β-party entanglement from those possible with β'-party entanglements, for $\beta \neq \beta'$.

In the Σ-QMAC no prior entanglement is allowed between the servers and the receiver Alice. Recall that entanglement between transmitter and receiver is required in the original setting where superdense coding is introduced [45]. This essential distinction also separates our work from prior efforts to classify entanglements according to their utility for DSC gains in [18], where the tasks chosen were simple communication tasks and prior entanglements between transmitters and receivers were allowed. In the Σ-QMAC setting since there is no prior entanglement between the servers and Alice, it follows from the Holevo bound that no DSC gain is possible for the direct communication task where each server wants to send an independent message to Alice. Thus, computation is essential to our setting which provides a richer space to explore multipartite entanglements.

In the Σ-QMAC the data streams may be replicated across multiple servers. The significance of this assumption is explained as follows. Without data-replication, the summation (Σ) is a *total* function of the computing parties' inputs, but with replication it is only a *partial* function. Arbitrarily large (e.g., exponential in the size of inputs) gains have been established for the exact computation of certain *partial* functions [20, 54] in the SMP model, i.e., without batch processing. The QMAC model, with batch processing, also allows arbitrarily large DSC gains for certain partial functions (see Appendix A). On the other hand, the largest observed DSC gain for exact computation of *total* functions thus far is only 2 in the SMP setting [23] with or without batch processing [18, 23, 36, 37]. To the best of our knowledge, DSC gains larger than 2 for exact computations of total functions, while unlikely, have not been formally ruled out. Specifically for our purpose, it is interesting that both directions are open for exact linear computations, i.e., there are no known instances of linear computations over the QMAC that achieve DSC gain larger than 2, nor is it known that gains larger than 2 are impossible in such settings.

The Σ-QMAC is related to the N-sum box [36], which is a black box abstraction of stabilizer based linear computations. The DSC gains in many previously studied settings, including recent applications in quantum private information retrieval (QPIR), can be realized through the N-sum box abstraction [18, 37–40, 46–49]. Nonetheless, it is important to note that the Σ-QMAC capacity formulation does not limit the coding schemes to the N-sum box, or to stabilizer based constructions in general. Indeed, the Σ-QMAC allows arbitrary entangled states to be shared among the servers, and stabilizer states are a very small fraction of those states. Similarly, the Σ-QMAC allows servers to perform arbitrary unitary transformations, whereas the N-sum box is limited to

X and Z gates, which represent a very small fraction of all possible unitaries. Whether the stabilizer based N-sum box suffices to achieve the capacity of the Σ-QMAC in all cases is another fundamental question to be answered in this work.

3 Overview of Contribution

The ability to precisely quantify *useful* multiparty entanglements via DSC gains over a QMAC boils down to following three requirements that must be simultaneously satisfied. 1) the desired computation must represent a natural task, 2) the exact capacity must be tractable, and 3) the capacity must in general require genuine multiparty entanglements, e.g., bipartite entanglements must not be sufficient to achieve the capacity in general. We show that the Σ-QMAC problem formulation satisfies all 3 criteria. Since sum-computation with distributed data is obviously a natural task, what remains is to show that the capacity is tractable and sensitive to genuine multiparty entanglement.

The tractability of capacity is established by an explicit capacity characterization in Theorem 1. The key ideas that make the capacity tractable include the capacity of sum-networks from network coding [28–32], the N-sum box abstraction [36], dual GRS codes [55], stabilizer based CSS code constructions [34, 35], and quantum-information theoretic converse arguments [56–58]. The capacity depends on both the data replication and entanglement distribution maps. For example, Corollary 4 shows that in the symmetric setting with $K = \binom{S}{\alpha}$ data-streams, each replicated among a distinct α-subset[4] of $[S]$, and $T = \binom{S}{\beta}$ quantum systems, each distributed among a distinct β-subset of $[S]$, the capacity of the Σ-QMAC is $\frac{1}{\beta T}\sum_{\gamma=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)} \min(\beta, 2\gamma) \cdot \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma}$.

For the motivating example in Fig. 1, let us explicitly state the capacity results for various entanglement distribution maps. It will be useful to adopt a more intuitive notation by denoting the servers $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4$ as $\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}$, respectively. The entanglement as shown in Fig. 1 can then be described as $(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$, reflecting the fact that there are 2 quantum systems, $\mathcal{Q}_1, \mathcal{Q}_2$, such that entangled subsystems of $\mathcal{Q}_1$ are distributed to the servers $\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}$, and entangled subsystems of $\mathcal{Q}_2$ are distributed to the servers $\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}$. The capacity for this, and various other entanglement distribution maps, is listed in Table 1.

[4] An α-subset of $[S]$ is a subset of $\{1, 2, \cdots, S\}$ that has cardinality α.

Entanglement distribution map	Capacity
$(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$	$4/5$
$(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}\}, \{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{d}}\}, \{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$	$3/4$
$(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}\}, \{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{d}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{d}}\}, \{\mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$	$3/4$
$(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$	$2/3$
$(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{d}}\}, \{\mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$	$2/3$
$(\{\mathcal{S}_{\mathsf{ab}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$	$2/3$
$(\{\mathcal{S}_{\mathsf{ab}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{d}}\}, \{\mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$	$2/3$
$(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}\}, \{\mathcal{S}_{\mathsf{d}}\})$	$1/2$
$(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}\}, \{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{bc}}\}, \{\mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}\}, \{\mathcal{S}_{\mathsf{d}}\})$	$1/2$
$(\{\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}\}, \{\mathcal{S}_{\mathsf{bc}}, \mathcal{S}_{\mathsf{d}}\})$	$1/2$
$(\{\mathcal{S}_{\mathsf{ab}}\}, \{\mathcal{S}_{\mathsf{ac}}\}, \{\mathcal{S}_{\mathsf{bc}}\}, \{\mathcal{S}_{\mathsf{d}}\})$	$2/5$

Table 1: Capacity of the Σ-QMAC in Fig. 1 for various entanglement distribution maps

Note that the last row of the table corresponds to the *unentangled* case, i.e., with no entanglements allowed between servers. The unentangled capacity of the Σ-QMAC for a given data replication map $\mathcal{W}$ is the same as the classical capacity for the corresponding setting. In this case, the unentangled capacity is $2/5$ computations/qudit, meaning that without quantum entanglement the fundamental limit dictates that each instance of the desired sum $\mathsf{A}+\mathsf{B}+\mathsf{C}+\mathsf{D}$ requires $5/2$ qudits ($5/2$ dits in the classical case) to be sent to Alice. The first row represents the opposite extreme, the fully entangled case that allows all 4 servers to be entangled, and we note that the capacity in this case is $4/5$ computations/qudit. For each specified entanglement distribution map the DSC gain is the ratio of the capacity for that case to the unentangled capacity. The DSC gain for the fully entangled setting is called the maximal distributed superdense coding gain. In this case, the maximal DSC gain is 2.

In terms of the main motivation of quantifying useful genuine multiparty entanglements via DSC gains, the capacity analysis allows us to draw various insights. Some of these are highlighted below along with pointers to relevant results in the paper.

1. The maximal DSC gain of Σ-QMAC[5] is 2.
2. The maximal DSC gain of 2 is achievable in the fully entangled Σ-QMAC if and only if the unentangled capacity is not more than $1/2$ (computations/qudit).
3. When the unentangled capacity is not less than $1/2$, the fully entangled Σ-QMAC has DSC gain exactly equal to the reciprocal of the unentangled capacity. The first three observations are implied by Corollary 3 of Section 5.
4. Bipartite (2-party) entanglement is in general insufficient to achieve the maximal DSC gain in the Σ-QMAC, even if unlimited bipartite entanglement is made available to every pair of transmitters. In other words, multiparty entanglement is *necessary* in general. This can be seen from Table 1 — the capacity is $2/3$ if only bipartite entanglement is allowed, whereas the capacity is $4/5$ with entanglement allowed across all four servers. The necessity of multiparty entanglements is also evident from Corollary 4 in Section 5, where explicit capacity expressions are provided for symmetric settings.

[5]The possibility of DSC gains larger than 2 for *vector* linear computations over the QMAC remains open.

5. If there is no data replication, i.e., each data-stream is only available to a unique server, then bipartite entanglement suffices to achieve the maximal DSC gain. This observation is based on Corollary 6 in Section 5.

6. The symmetric setting in Corollary 4 also reveals that both extremes of too much data replication and too little data replication require relatively little entanglement to achieve their maximal DSC gain, rather the intermediate regimes of data replication are the ones that require the most entanglement to achieve their maximal DSC gain. See discussion in Section 6.3.

7. The minimal entanglement-size β such that β-party entanglement is necessary to achieve a desired (feasible) DSC gain value in a Σ-QMAC can be determined from the capacity characterizations. For example, from the capacity characterizations in Table 1 we note that the maximal DSC gain cannot be achieved with 3-party entanglement, i.e., 4-party entanglement is necessary to achieve maximal DSC gain. On the other hand, the best DSC gain with 3-party entanglements is only $(3/4)/(2/5) = 15/8$ which can also be achieved with 2-party entanglements, i.e., the minimal entanglement-size needed for DSC gain $15/8$ is $\beta = 2$.

8. 3-party entanglement is never *necessary* to achieve the capacity of the Σ-QMAC. Any 3-party entanglement can be replaced by 2-party entanglements with the capacity unchanged. This observation is based on Corollary 7 of Section 5. Some discussion with an example is presented in Section 6.5.

9. For every $S \neq 3$, there is a Σ-QMAC setting with S servers where S-party entanglement is *necessary* to achieve the maximal DSC gain. In such settings, even with unlimited $S-1$ party entanglements among all $(S-1)$-subsets of servers, the DSC gain is strictly smaller than that with S-party entanglement. This observation is based on Corollary 8 of Section 5. Related discussion is provided in Section 6.6.

10. Entanglements restricted to stabilizer states, along with Pauli operations (X and Z gates) at the servers (i.e., coding via the N-sum box abstraction) suffice to achieve the capacity of the Σ-QMAC. Thus, while the capacity formulation does allow non-stabilizer states and general unitary operations, neither of those can improve the DSC gains in the Σ-QMAC. This observation emerges from the proof of achievability of Theorem 1.

Notation: $\mathbb{N}$ denotes the set of positive integers. $\mathbb{Z}^+ = \{0\} \cup \mathbb{N}$. For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, \cdots, n\}$. For $n_1, n_2 \in \mathbb{N}$, $[n_1 : n_2]$ denotes the set $\{n_1, n_1+1, \cdots, n_2\}$ if $n_1 \leq n_2$ and $\emptyset$ otherwise. For a set $\mathcal{X}$, define $\mathcal{X}^n \triangleq \mathcal{X} \times \mathcal{X} \times \cdots \times \mathcal{X}$ as the n-fold Cartesian product. Define compact notations $A^{[n]} \triangleq (A^{(1)}, A^{(2)}, \cdots, A^{(n)})$ and $A_{[n]} \triangleq (A_1, A_2, \cdots, A_n)$. $\mathbb{F}_d$ denotes the finite field with $d = p^r$ a power of a prime. $\mathbb{R}$ denotes the set of real numbers. $\mathbb{R}_+$ denotes the set of non-negative real numbers. $\mathbb{C}$ denotes the set of complex numbers. For a field $\mathbb{F}$, $\mathbb{F}^{a\times b}$ denotes the set of $a \times b$ matrices with elements in $\mathbb{F}$. $\mathsf{tr}(M)$ denotes the trace of a matrix M. For a matrix M with elements in $\mathbb{C}$, $M^\dagger$ denotes its conjugate transpose. $\mathbf{I}_a$ denotes the $a \times a$ identity matrix. $\mathbf{0}_{a\times b}$ denotes the zero matrix with size $a \times b$. $\mathsf{Pr}(E)$ denotes the probability of an event E. $\mathsf{Pr}(E_1|E_2)$ denotes the conditional probability of E_1 given E_2. $(x)^+ \triangleq \max(x, 0)$. For $m, n \in \mathbb{Z}^+, m \leq n$, $\binom{n}{m} \triangleq \frac{n!}{m!(n-m)!}$ denotes the binomial coefficient. For a set $\mathcal{N}$, the set of its cardinality-m sub-sets is denoted as $\binom{\mathcal{N}}{m} \triangleq \{\mathcal{A} \subset \mathcal{N} \mid |\mathcal{A}| = m\}$ if $|\mathcal{N}| \geq m$. The notation $2^{\mathcal{N}}$ denotes the power set of $\mathcal{N}$. The notation $f : \mathcal{A} \mapsto \mathcal{B}$ denotes a map f from $\mathcal{A}$ to $\mathcal{B}$. If f is a bijection from $\mathcal{A}$ to $\mathcal{B}$, we write

$f : \mathcal{A} \leftrightarrow \mathcal{B}$ and denote the inverse of f as f^{-1}. The dimension of a quantum system $\mathcal{Q}$ is denoted as $|\mathcal{Q}|$.

4 Problem Formulation

4.1 Σ-QMAC

The Σ-QMAC problem is specified by a 6-tuple $\big(\mathbb{F}_d, S, K, T, \mathcal{W}, \mathcal{E}\big)$. $\mathbb{F}_d$ is a finite field of order d with $d = p^r$ being a power of a prime. S is the number of servers. K is the number of independent classical data-streams, denoted as $\mathsf{W}_1, \mathsf{W}_2, \cdots, \mathsf{W}_K$. The k^{th} data stream, W_k, is comprised of symbols $\mathsf{W}_k^{(\ell)} \in \mathbb{F}_d, \ell \in \mathbb{N}$. T is the number of independent quantum resources, denoted as $\mathcal{Q}_1, \mathcal{Q}_2, \cdots, \mathcal{Q}_T$. The *data replication map* is a mapping $\mathcal{W} : [K] \to 2^{[S]}$ that identifies $\mathcal{W}(k) \subset [S]$ as the subset of servers where W_k is available. The *entanglement distribution map* is a mapping $\mathcal{E} : [T] \to 2^{[S]}$ that identifies $\mathcal{E}(t) \subset [S]$ as the subset of servers among which the quantum system $\mathcal{Q}_t$ is distributed. Such a subset of servers is referred to as a *clique* in this paper. The quantum system $\mathcal{Q}_t$ is partitioned into entangled subsystems $\mathcal{Q}_{t,s}, s \in \mathcal{E}(t)$ such that Server s receives the quantum subsystem $\mathcal{Q}_{t,s}$ from the quantum system $\mathcal{Q}_t = (\mathcal{Q}_{t,s}, s \in \mathcal{E}(t))$.

4.2 Feasible Quantum Coding Schemes

A quantum coding scheme is specified by a 6-tuple

$$\big(L,\ ((\delta_{t,s})_{s\in\mathcal{E}(t)})_{t\in[T]},\ \rho_{[T]},\ ((\Phi_{t,s})_{s\in\mathcal{E}(t)})_{t\in[T]},\ (\{M_{t,y}\}_{y\in\mathcal{Y}_t})_{t\in[T]},\ \Psi\big). \tag{1}$$

$L \in \mathbb{N}$ is the batch size, which is the number of sums to be computed by the coding scheme, i.e., the coding scheme allows Alice to compute $\mathsf{W}_\Sigma(\ell)$ for all $\ell \in [L]$. For $k \in [K]$, denote the first L instances of the k^{th} data stream as $\mathsf{W}_k^{[L]} = \big(\mathsf{W}_k^{(1)}, \mathsf{W}_k^{(2)}, \cdots, \mathsf{W}_k^{(L)}\big) \in \mathbb{F}_d^L$, and the desired computation at Alice as $\mathsf{W}_\Sigma^{[L]} = \big(\mathsf{W}_\Sigma^{(1)}, \mathsf{W}_\Sigma^{(2)}, \cdots, \mathsf{W}_\Sigma^{(L)}\big) \in \mathbb{F}_d^L$, where $\mathsf{W}_\Sigma(\ell) \triangleq \sum_{k=1}^K \mathsf{W}_k(\ell), \forall \ell \in [L]$. For $t \in [T]$ and $s \in \mathcal{E}(t)$, $\delta_{t,s} \in \mathbb{Z}^+$ specifies the dimension of the quantum subsystem $\mathcal{Q}_{t,s}$, i.e., $|\mathcal{Q}_{t,s}| = \delta_{t,s}$. For $t \in [T]$, say the t^{th} clique is $\mathcal{E}(t) = \{s_1, s_2, \cdots, s_{|\mathcal{E}(t)|}\}$. The quantum system $\mathcal{Q}_t = \mathcal{Q}_{t,s_1}\mathcal{Q}_{t,s_2}\cdots\mathcal{Q}_{t,s_{|\mathcal{E}(t)|}}$ is prepared in the initial state $\rho_t \in \mathbb{C}^{|\mathcal{Q}_t|\times|\mathcal{Q}_t|}$. $\mathcal{Q}_1, \mathcal{Q}_2, \cdots, \mathcal{Q}_T$ are unentangled with each other. Without loss of generality, we assume that the initial state of the composite system is a pure state, and thus it can be written as $\rho = \rho_1 \otimes \rho_2 \otimes \cdots \otimes \rho_T$. For $t \in [T]$, Server $s \in \mathcal{E}(t)$ applies a unitary operator $U_{t,s} = \Phi_{t,s}(\mathsf{W}_k^{[L]}, k : s \in \mathcal{W}(k))$ to $\mathcal{Q}_{t,s}$. Thus, the resulting state of $\mathcal{Q}_t$ is $\rho_t' = U_t \rho_t U_t^\dagger$, where $U_t \triangleq U_{t,s_1} \otimes U_{t,s_2} \otimes \cdots \otimes U_{t,s_{|\mathcal{E}(t)|}}$. Note that since $\mathcal{Q}_1, \mathcal{Q}_2, \cdots, \mathcal{Q}_T$ are unentangled initially, and separate operations are done for $\mathcal{Q}_1, \mathcal{Q}_2, \cdots, \mathcal{Q}_T$, they remain unentangled after the servers apply the operations. All subsystems are then sent to Alice, who performs separate quantum measurements (POVM) on each of the T quantum systems. Specifically, for $t \in [T]$, the set of operators for the measurement of $\mathcal{Q}_t$ is specified as $\{M_{t,y}\}_{y\in\mathcal{Y}_t}$ by the coding scheme. The output of the measurement is denoted as Y_t, which is a random variable with realizations in $\mathcal{Y}_t$. Finally, the function $\Psi : \mathcal{Y}_1 \times \mathcal{Y}_2 \cdots \times \mathcal{Y}_T \to \mathbb{F}_d^{L\times 1}$ maps the outputs of the measurements $Y_{[T]} = (Y_1, Y_2, \cdots, Y_T)$ to the desired computation (sum), i.e., $\mathsf{W}_\Sigma^{[L]} = \Psi(Y_1, \cdots, Y_T)$. Any feasible coding scheme must work for all d^{KL} realizations of $\big(\mathsf{W}_1^{[L]}, \mathsf{W}_2^{[L]}, \cdots, \mathsf{W}_K^{[L]}\big)$. Let $\mathfrak{C}$ denote the set of such coding schemes.

4.3 Feasible Region and Capacity

For the Σ-QMAC$\big(\mathbb{F}_d, S, K, T, \mathcal{W}, \mathcal{E}\big)$, the download-cost per computation (qudits/dit) tuple,

$$\boldsymbol{\Delta} = (\Delta_{t,s})_{t\in[T],s\in\mathcal{E}(t)} \in \mathbb{R}_+^{\Gamma},\ \ \Gamma \triangleq \sum_{t\in[T]} |\mathcal{E}(t)|, \tag{2}$$

is said to be feasible, if there exists a coding scheme

$$\big(L,\ ((\delta_{t,s})_{s\in\mathcal{E}(t)})_{t\in[T]},\ \rho_{[T]},\ ((\Phi_{t,s})_{s\in\mathcal{E}(t)})_{t\in[T]},\ (\{M_{t,y}\}_{y\in\mathcal{Y}_t})_{t\in[T]},\ \Psi\big) \in \mathfrak{C}$$

such that

$$\Delta_{t,s} \geq \frac{\log_d |\mathcal{Q}_{t,s}|}{L} = \frac{\log_d \delta_{t,s}}{L}, \qquad\qquad \forall t \in [T], s \in \mathcal{E}(t). \tag{3}$$

Define $\mathcal{D}$ as the closure of the set of all feasible download-cost tuples $\boldsymbol{\Delta}$ so that any $\boldsymbol{\Delta}$ inside $\mathcal{D}$ is feasible, and any $\boldsymbol{\Delta}$ outside $\mathcal{D}$ is not feasible. In terms of computation rates (dits of computation/qudit of download), a rate R is feasible if there exists a coding scheme in $\mathfrak{C}$ such that

$$R \leq \frac{L}{\sum_{t\in[T],s\in\mathcal{E}(t)} \log_d |\mathcal{Q}_{t,s}|} = \frac{L}{\sum_{t\in[T],s\in\mathcal{E}(t)} \log_d \delta_{t,s}}. \tag{4}$$

Define

$$C \triangleq \sup_{\mathfrak{C}} R \tag{5}$$

as the computation capacity. Note that a capacity C characterization is implied by a characterization of $\mathcal{D}$ because $C = (\min_{\Delta\in\mathcal{D}} \sum_{t\in[T],s\in\mathcal{E}(t)} \Delta_{t,s})^{-1}$. Since S, K, T can be inferred from $\mathcal{W}, \mathcal{E}$, a Σ-QMAC problem is fully specified by the parameters $(\mathbb{F}_d, \mathcal{W}, \mathcal{E})$. As our first result (Theorem 1) will show, the capacity is independent of $\mathbb{F}_d$ (which reflects the merit of using qudit to measure the cost). Therefore the capacity $C(\mathcal{W}, \mathcal{E})$ is only a function of $(\mathcal{W}, \mathcal{E})$. While comparing capacities of problems that have the same data replication map but different entanglement distribution maps, the data replication map $\mathcal{W}$ may be occasionally omitted for brevity when it is clear from the context.

4.4 Fully-entangled, Fully-unentangled and Fully-β-party-entangled Capacities

Symmetric entanglement distribution maps can be especially insightful because the symmetry facilitates compact capacity descriptions that are easier to compare. To prepare for discussions of symmetric entanglements, let us define the *fully-β-party-entangled* setting as the entanglement distribution map comprised of all cliques of size β. Note that this allows genuine multiparty entanglement to be established among every subset of servers containing no more than β servers. Formally, *fully-β-party-entangled* setting[6] refers to the *bijective* entanglement distribution map,

$$\mathcal{E}^{(\beta)} : [\binom{S}{\beta}] \leftrightarrow \binom{[S]}{\beta}. \tag{6}$$

[6]We may shorten 'fully-β-party-entangled' to simply 'β-party-entangled' in subsequent discussions when the context is obvious.

For example, with $S = 4$ servers, fully-2-party-entangled setting means that $\binom{4}{2} = 6$ separate quantum systems are available, each comprised of 2 entangled subsystems that are distributed among a distinct pair (2-clique) of servers. $\mathcal{E}^{(2)}$ is a bijection from $\{1, 2, \cdots, 6\}$ to the set of all 2-subsets of $[4]$, i.e., $\{\{1,2\},\{2,3\},\{3,4\},\{4,1\},\{1,3\},\{2,4\}\}$. For example, $\mathcal{E}^{(2)}(1) = \{1,2\}$, $\mathcal{E}^{(2)}(2) = \{2,3\}$, $\mathcal{E}^{(2)}(3) = \{3,4\}$, $\mathcal{E}^{(2)}(4) = \{4,1\}$, $\mathcal{E}^{(2)}(5) = \{1,3\}$, $\mathcal{E}^{(2)}(6) = \{2,4\}$). Given a data replication map $\mathcal{W}$, the fully-β-party-entangled capacity is correspondingly defined as,

$$C^{(\beta)}(\mathcal{W}) \triangleq C(\mathcal{W}, \mathcal{E}^{(\beta)}). \tag{7}$$

Extreme cases of fully-β-party-entangled capacity include the $\beta = S$ setting, known simply as the fully-entangled capacity $C^{\text{fullent}}(\mathcal{W}) = C^{(S)}(\mathcal{W})$, and the $\beta = 1$ setting, known simply as the *fully-unentangled* capacity $C^{\text{unent}}(\mathcal{W}) = C^{(1)}(\mathcal{W})$.

4.5 Distributed Superdense Coding (DSC) Gain

Given data replication map $\mathcal{W}$ and entanglement distribution map $\mathcal{E}$, the distributed super dense coding (DSC) gain is defined as the ratio

$$C(\mathcal{W}, \mathcal{E})/C^{\text{unent}}(\mathcal{W}), \tag{8}$$

which indicates the multiplicative gain, compared to the fully-unentangled capacity $C^{\text{unent}}(\mathcal{W})$, that is enabled by quantum entanglement subject to the entanglement distribution map $\mathcal{E}$. The *maximal* DSC gain for data replication map $\mathcal{W}$ is defined as $C^{\text{fullent}}(\mathcal{W})/C^{\text{unent}}(\mathcal{W})$, which is the ratio of the fully-entangled capacity to the fully-unentangled capacity.

5 Results

The exact capacity of the Σ-QMAC is fully characterized in the following theorem.

Theorem 1 (Σ-QMAC)**.** *The capacity of the Σ-QMAC $\left(\mathbb{F}_d, S, K, T, \mathcal{W}, \mathcal{E}\right)$, is*

$$C(\mathcal{W}, \mathcal{E}) = \left(\min_{\mathbf{\Delta} \in \mathcal{D}(\mathcal{W}, \mathcal{E})} \sum_{t \in [T], s \in \mathcal{E}(t)} \Delta_{t,s} \right)^{-1} \tag{9}$$

with the feasible download-cost region characterized as ($\Gamma \triangleq \sum_{t \in [T]} |\mathcal{E}(t)|$),

$$\mathcal{D}(\mathcal{W}, \mathcal{E}) = \left\{ \mathbf{\Delta} \in \mathbb{R}_+^{\Gamma} \;\middle|\; \sum_{t \in [T]} \min \left\{ \sum_{s \in \mathcal{E}(t)} \Delta_{t,s}, \sum_{s \in \mathcal{E}(t) \cap \mathcal{W}(k)} 2\Delta_{t,s} \right\} \geq 1, \forall k \in [K] \right\}. \tag{10}$$

The proof of Theorem 1 appears in Sections 7 and 8. Note that Theorem 1 characterizes the capacity of the Σ-QMAC in terms of the solution of a linear program that finds the minimum download cost over the feasible region $\mathcal{D}$ that is explicitly characterized. The capacity does not depend on the field $\mathbb{F}_d$. In fact the capacity depends *only* on the data replication and entanglement distribution maps $(\mathcal{W}, \mathcal{E})$ since the remaining parameters S, K, T can be inferred from $\mathcal{W}, \mathcal{E}$.

The remainder of this section identifies a few interesting corollaries that follow from Theorem 1. We start with the specializations for the opposite extremes, fully-entangled and fully-unentangled capacities, stated as corollaries next.

Corollary 1 (Fully-entangled)**.** *The fully-entangled capacity for a data replication map $\mathcal{W}$ is,*

$$C^{\text{fullent}}(\mathcal{W}) = \left(\min_{(\Delta_1,\cdots,\Delta_S)\in\mathcal{D}^{\text{fullent}}(\mathcal{W})} \sum_{s\in[S]} \Delta_s \right)^{-1}, \tag{11}$$

where

$$\mathcal{D}^{\text{fullent}}(\mathcal{W}) = \left\{ (\Delta_1,\cdots,\Delta_S) \in \mathbb{R}_+^S \,\middle|\, \sum_{s\in[S]} \Delta_s \geq 1, \sum_{s\in\mathcal{W}(k)} \Delta_s \geq 1/2, \forall k \in [K] \right\}. \tag{12}$$

Corollary 2 (Fully-unentangled)**.** *The fully-unentangled capacity for a data replication map $\mathcal{W}$ is,*

$$C^{\text{unent}}(\mathcal{W}) = \left(\min_{(\Delta_1,\cdots,\Delta_S)\in\mathcal{D}^{\text{unent}}(\mathcal{W})} \sum_{s\in[S]} \Delta_s \right)^{-1} \tag{13}$$

where

$$\mathcal{D}^{\text{unent}}(\mathcal{W}) = \left\{ (\Delta_1,\cdots,\Delta_S) \in \mathbb{R}_+^S \;\middle|\; \sum\nolimits_{s\in\mathcal{W}(k)} \Delta_s \geq 1, \forall k \in [K] \right\}. \tag{14}$$

From Corollary 1 and Corollary 2 we have the following characterization of the maximal DSC gain for any data replication map $\mathcal{W}$.

Corollary 3 (Maximal DSC gain)**.** *The maximal distributed superdense coding gain for the data replication map $\mathcal{W}$ is,*

$$C^{\text{fullent}}(\mathcal{W})/C^{\text{unent}}(\mathcal{W}) = \min\left(2, 1/C^{\text{unent}}(\mathcal{W})\right). \tag{15}$$

The proof of Corollary 3 is relegated to Appendix B.

Next we explore a class of Σ-QMAC settings with symmetric data replication and entanglement distribution maps. A symmetric setting is specified by three parameters, S, α and β. The data replication map is fully symmetric, so that for each α-subset of $[S]$ there is a unique data stream replicated among this subset of servers. The goal is to characterize explicitly the fully-β-party-entangled capacity for such data replication maps. The explicit characterization is provided next.

Corollary 4 (Symmetric)**.** *If $\alpha \in [S]$, $K = \binom{S}{\alpha}$, $T = \binom{S}{\beta}$, and $\mathcal{W} : [K] \leftrightarrow \binom{[S]}{\alpha}$ and $\mathcal{E} : [T] \leftrightarrow \binom{[S]}{\beta}$ are bijective mappings, then the Σ-QMAC capacity (denoted as $C_\alpha^{(\beta)}$) is*

$$C_\alpha^{(\beta)} = \frac{1}{\beta T} \sum_{\gamma=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)} \min(\beta, 2\gamma) \cdot \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma} \tag{16}$$

$$= \frac{2\alpha}{S} - \frac{1}{\beta T} \sum_{\gamma=\max(\alpha+\beta-S,\lceil\beta/2\rceil)}^{\min(\alpha,\beta)} (2\gamma-\beta) \cdot \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma} \tag{17}$$

$$= 1 - \frac{1}{\beta T} \sum_{\gamma=(\alpha+\beta-S)^+}^{\min(\alpha,\lfloor\beta/2\rfloor)} (\beta-2\gamma) \cdot \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma} \tag{18}$$

The proof of Corollary 4 is relegated to Appendix C.

The next corollary sheds light on $C^{(2)}$, i.e., the capacity with only bipartite entanglements. Let us first provide the necessary context before presenting the next corollary. Given a Σ-QMAC problem $\mathcal{P}$ with data replication $\mathcal{W}$, S servers and K data streams, we want to explicitly characterize the 2-party-entangled capacity $C^{(2)}(\mathcal{W})$. To do so, let us construct another hypothetical Σ-QMAC problem referred to as $\widetilde{\mathcal{P}}$ with data replication map $\widetilde{\mathcal{W}}$, $\binom{S}{2}$ servers and the same K data streams as in $\mathcal{P}$. Now let us specify $\widetilde{\mathcal{W}}$. Each server in $\widetilde{\mathcal{P}}$ is indexed by a 2-element set $\{i,j\} \subset [S]$. Server $\mathcal{S}_{\{i,j\}}$ in $\widetilde{\mathcal{P}}$ has the access to the data streams that are available to Servers $\mathcal{S}_i$ or $\mathcal{S}_j$ in $\mathcal{P}$. In other words, a data stream is available to Server $\mathcal{S}_{\{i,j\}}$ in $\widetilde{\mathcal{P}}$ if and only if that data stream is available to either Server $\mathcal{S}_i$ or Server $\mathcal{S}_j$ (or to both) in $\mathcal{P}$. Mathematically, $\widetilde{\mathcal{W}}(k) = \left\{\{i,j\} \in \binom{[S]}{2} \mid \left(\{i,j\} \cap \mathcal{W}(k)\right) \neq \emptyset\right\}$ for $k \in [K]$. Now we are ready to present the next corollary.

Corollary 5 (Fully-2-party-entangled capacity)**.** *$C^{(2)}(\mathcal{W}) = C^{\text{unent}}(\widetilde{\mathcal{W}})$. In addition, it can be shown that $C^{(2)}(\mathcal{W})$ can always be achieved by a scheme that involves only the 2-sum protocol.*[7]

The proof of Corollary 5 is relegated to Appendix D. In plain words, Corollary 5 states that the 2-party-entangled capacity for the data replication map $\mathcal{W}$ is equal to the fully-unentangled capacity for the data replication map $\widetilde{\mathcal{W}}$, comprised of a new set of servers that are obtained by merging pairs of original servers.

Corollary 6 (Disjoint data)**.** *Given data replication map $\mathcal{W}$ with S servers and K data streams, if $S \geq 2$ and each data-stream is available to only one server, i.e., $|\mathcal{W}(k)| = 1$ for all $k \in [K]$, then $C^{\text{fullent}}(\mathcal{W}) = C^{(2)}(\mathcal{W}) = 2/S$.*

In other words, if no data stream is replicated across more than one server, then genuine multi-party entanglement (between more than 2 parties) is not needed, i.e., the fully-entangled capacity is equal to the 2-party-entangled capacity. Together with Corollary 5, this implies that 2-sum protocol based schemes suffice to achieve the fully-entangled capacity in this case.

Proof. Since each server can locally add the data streams and regard the sum as one data stream, it suffices to consider $K = S$ data streams such that each server has a unique data stream. The reduced setting belongs to the symmetric settings specified in Corollary 4 with S servers and $\alpha = 1$. It can then be verified by (17) that $C_1^{(2)} = C_1^{(S)} = 2/S$. □

Corollary 7 (3-party entanglement is unnecessary)**.** *Given any Σ-QMAC problem with data replication map $\mathcal{W}$ and entanglement distribution map $\mathcal{E}$ that identifies a 3-party clique, say $\mathcal{E}(t) = \{s_1, s_2, s_3\}$ for some $t \in [T]$, consider another entanglement distribution map $\mathcal{E}'$, which is created by first making a copy of $\mathcal{E}$, and then replacing the 3-party clique $\{s_1, s_2, s_3\}$ with three 2-party cliques $\{s_1, s_2\}, \{s_1, s_3\}$ and $\{s_2, s_3\}$. Then we always have $C(\mathcal{W}, \mathcal{E}) = C(\mathcal{W}, \mathcal{E}')$.*

The proof of Corollary 7 is relegated to Appendix E. It shows that any 3-party entanglement can be substituted by 2-party entanglements established by the same three servers, for the purpose of Σ-QMAC capacity (not necessarily for other function computations). Note that it immediately follows that $C^{(3)}(\mathcal{W}) = C^{(2)}(\mathcal{W})$ for any data replication map $\mathcal{W}$.

[7] Schemes that only apply the 2-sum protocol are special cases of the quantum coding schemes formulated in Section 4.2 when $T = \binom{S}{2}$ and $\mathcal{E} : [T] \leftrightarrow \binom{[S]}{2}$. In other words, each of the T quantum systems $\mathcal{Q}_1, \mathcal{Q}_2, \cdots, \mathcal{Q}_T$ is available to a unique pair of the S servers.

Corollary 8 (Necessity of multiparty entanglements)**.** *For any $S \geq 2$ and $S \neq 3$, there exists a data replication map $\mathcal{W}$ with S servers for which $C^{(S)}(\mathcal{W}) > C^{(S-1)}(\mathcal{W})$.*

The proof of Corollary 8 is relegated to Appendix F. The corollary states that given the number of servers S, if $S \geq 2$ and $S \neq 3$, there is a data replication map $\mathcal{W}$ with S servers for which the fully-entangled capacity $C^{\text{fullent}}(\mathcal{W})$ is strictly greater than the $(S-1)$-party-entangled capacity $C^{(S-1)}(\mathcal{W})$. In other words, in order to attain the maximal DSC gain for the Σ-QMAC with this data replication map, one must allow the entanglement to be established among all S servers. A concrete example for this corollary can be found in Section 6.6.

Finally, let us explore the *separability* of Σ-QMAC settings, explained as follows. Consider two Σ-QMAC problems, labeled Σ-QMAC$_1$ and Σ-QMAC$_2$. Say Σ-QMAC$_i$, $i \in \{1,2\}$, has $S_i > 0$ servers and $K_i > 0$ messages. It will be convenient to label the K_1 messages of Σ-QMAC$_1$ with indices $[1 : K_1]$, while the K_2 messages of Σ-QMAC$_1$ are assigned indices $[K_1 + 1 : K_1 + K_2]$. Similarly, the S_1 servers of Σ-QMAC$_1$ have indices $[1 : S_1]$, while the S_2 servers of Σ-QMAC$_2$ have indices $[S_1 + 1 : S_1 + S_2]$. The data replication maps for Σ-QMAC$_1$ and Σ-QMAC$_2$ are $\mathcal{W}_1$ and $\mathcal{W}_2$, respectively. Combining Σ-QMAC$_1$ and Σ-QMAC$_2$, let us construct Σ-QMAC$_3$, with $S_3 = S_1 + S_2$ servers, $K_3 = K_1 + K_2$ messages, and the data replication map,

$$\mathcal{W}_3(k) = \begin{cases} \mathcal{W}_1(k), & k \in [1 : K_1], \\ \mathcal{W}_2(k), & k \in [K_1 + 1 : K_1 + K_2]. \end{cases} \tag{19}$$

Let $C^{\text{fullent}}(\mathcal{W}_i)$ denote the fully entangled capacity of Σ-QMAC$_i$, where the entanglement is limited to the S_i servers of Σ-QMAC$_i$, $i \in \{1,2,3\}$. Now we are ready to formalize the notion of separability. We say that Σ-QMAC$_1$ and Σ-QMAC$_2$ are *separable* iff,

$$1/C^{\text{fullent}}(\mathcal{W}_3) = 1/C^{\text{fullent}}(\mathcal{W}_1) + 1/C^{\text{fullent}}(\mathcal{W}_2). \tag{20}$$

Recall that the reciprocals of capacities are the optimal communication costs per computation. Note that if Σ-QMAC$_1$ and Σ-QMAC$_2$ are *separable* then it suffices to have separate entanglements established among Servers $[1 : S_1]$ and Servers $[S_1 + 1 : S_1 + S_2]$ to achieve the maximal DSC gain for Σ-QMAC$_3$. The following corollary specifies a necessary and sufficient condition for separability.

Corollary 9 (Separability)**.** *Σ-QMAC$_1$ and Σ-QMAC$_2$ are separable if and only if they each have DSC gain of 2. Formally,*

$$\left.\begin{array}{l} C^{\text{fullent}}(\mathcal{W}_1)/C^{\text{unent}}(\mathcal{W}_1) = 2 \\ C^{\text{fullent}}(\mathcal{W}_2)/C^{\text{unent}}(\mathcal{W}_2) = 2 \end{array}\right\} \iff 1/C^{\text{fullent}}(\mathcal{W}_3) = 1/C^{\text{fullent}}(\mathcal{W}_1) + 1/C^{\text{fullent}}(\mathcal{W}_2). \tag{21}$$

Proof. Let us show that Σ-QMAC$_3$ has DSC gain of 2, regardless of the DSC gain of Σ-QMAC$_1$ and Σ-QMAC$_2$. This is because by definition Σ-QMAC$_3$ is composed of two separate problems with independent datasets and two separate sets of servers that have these datasets, which, together with Corollary 2 implies that $1/C_3^{\text{unent}}(\mathcal{W}) = 1/C_1^{\text{unent}}(\mathcal{W}) + 1/C_2^{\text{unent}}(\mathcal{W})$. It is also readily verified that $1/C^{\text{unent}}(\mathcal{W}_1) \geq 1$ and $1/C^{\text{unent}}(\mathcal{W}_2) \geq 1$ by Corollary 2. Therefore, we have $1/C^{\text{unent}}(\mathcal{W}_3) \geq 2$. It then follows from Corollary 3 that $C^{\text{fullent}}(\mathcal{W}_3)/C^{\text{unent}}(\mathcal{W}_3) = \min\{2, 1/C^{\text{unent}}(\mathcal{W}_3)\} = 2$.

To show the forward direction ' $\implies$ ' in (21), suppose we are given that Σ-QMAC$_1$ and Σ-QMAC$_2$ each have DSC gain of 2. Since all three problems have DSC gain 2, we have that $1/C^{\text{fullent}}(\mathcal{W}_3) = 0.5/C^{\text{unent}}(\mathcal{W}_3) = 0.5/C^{\text{unent}}(\mathcal{W}_1) + 0.5/C^{\text{unent}}(\mathcal{W}_2) = 1/C^{\text{fullent}}(\mathcal{W}_1) + 1/C^{\text{fullent}}(\mathcal{W}_2)$, which proves the forward direction of (21).

For the reverse direction, we are given that $1/C^{\text{fullent}}(\mathcal{W}_3) = 1/C^{\text{fullent}}(\mathcal{W}_1) + 1/C^{\text{fullent}}(\mathcal{W}_2)$. To set up a proof by contradiction, suppose without loss of generality, that $C^{\text{fullent}}(\mathcal{W}_1)/C^{\text{unent}}(\mathcal{W}_1) < 2$. Then we have $1/C^{\text{fullent}}(\mathcal{W}_3) = 0.5/C^{\text{unent}}(\mathcal{W}_3) = 0.5/C^{\text{unent}}(\mathcal{W}_1) + 0.5/C^{\text{unent}}(\mathcal{W}_2) < 1/C^{\text{fullent}}(\mathcal{W}_1) + 1/C^{\text{fullent}}(\mathcal{W}_2)$. The contradiction completes the proof. □

6 Examples

In this section we present examples to illustrate the results.

6.1 Example: Achieving Fully-Entangled Capacity C^{fullent} for the Data Replication Map of Fig. 1

Consider the data replication map $\mathcal{W}$ as in Fig. 1. Let us sketch the solution for the fully-entangled setting, i.e., $\mathcal{E} = \{\{1, 2, 3, 4\}\}$. For intuitive notation, let us use subscripts ab, ac, bc, d to represent 1, 2, 3, 4, respectively, reflecting the data-streams available at the corresponding servers. For example, we indicate Server $\mathcal{S}_1$ as $\mathcal{S}_{\mathsf{ab}}$, making it explicit that this server has data-streams A, B. With this notation, the feasible region in Corollary 1 is,

$$\mathcal{D}^* = \left\{ (\Delta_{\mathsf{ab}}, \Delta_{\mathsf{ac}}, \Delta_{\mathsf{bc}}, \Delta_{\mathsf{d}}) \in \mathbb{R}_+^4 \;\middle|\; \begin{array}{l} \Delta_{\mathsf{ab}} + \Delta_{\mathsf{ac}} + \Delta_{\mathsf{bc}} + \Delta_{\mathsf{d}} \geq 1, \\ \Delta_{\mathsf{ab}} + \Delta_{\mathsf{ac}} \geq 1/2, \\ \Delta_{\mathsf{ab}} + \Delta_{\mathsf{bc}} \geq 1/2, \\ \Delta_{\mathsf{ac}} + \Delta_{\mathsf{bc}} \geq 1/2, \\ \Delta_{\mathsf{d}} \geq 1/2. \end{array} \right\}. \tag{22}$$

From the converse standpoint, let us note informally that of the 5 bounds that appear in (22), the first bound says that the normalized *total* download cost is at least 1 qudit/dit. This is because there is no entanglement between the servers and Alice, so it follows from the Holevo bound [56] that one qudit cannot carry more than one dit of information. The remaining four bounds are typical cut-set arguments, by separating the parties (servers and Alice) into two groups such that the servers that know one of the data streams are collectively regarded as the transmitter, while the other servers join Alice as the receiver. Since now entanglement can be established between the transmitter and the receiver, it follows from (e.g., [18, 59]) that the one qudit can carry at most 2 dits of information, yielding the factor $1/2$ on the RHS of the other four bounds.

Minimizing $\Delta_{\mathsf{ab}} + \Delta_{\mathsf{ac}} + \Delta_{\mathsf{bc}} + \Delta_{\mathsf{d}}$ over $\mathcal{D}$ leads to a linear program with optimal value $5/4$, thus establishing the fully-entangled capacity for this example as $C^{\text{fullent}} = 4/5$. To show the achievability of $4/5$, we specify a coding scheme that allows Alice to recover $L = 4$ instances of the desired sums, based on an $(N = 5)$-sum box in $\mathbb{F}_d$ so that in the 5-sum box Server $\mathcal{S}_{\mathsf{ab}}$ controls 1 pair of inputs x_1, x_6; Server $\mathcal{S}_{\mathsf{ac}}$ controls 1 pair of inputs x_2, x_7; Server $\mathcal{S}_{\mathsf{bc}}$ controls 1 pair of inputs x_3, x_8; and Server $\mathcal{S}_{\mathsf{d}}$ controls 2 pairs of inputs x_4, x_5, x_9, x_{10}. The input-output relationship for the 5-sum box is $\mathbf{y} = \mathbf{M}\mathbf{x}$ with the transfer function $\mathbf{M} \in \mathbb{F}_d^{5\times 10}$ specified as,

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}. \tag{23}$$

It is easily verified that $\mathbf{M}\mathbf{J}_{10}\mathbf{M}^T = \mathbf{0}_{5\times 5}$ and thus it is a valid 5-sum box.[8]

To the output $\mathbf{y} \in \mathbb{F}_d^{5\times 1}$, Alice applies a 4×5 decoding matrix V_{dec} specified as,

$$V_{\text{dec}} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}. \tag{24}$$

The input vector $\mathbf{x} \in \mathbb{F}_d^{10\times 1}$ is specified as,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{10} \end{bmatrix} = \begin{bmatrix} V^{\mathsf{a}}_{\mathsf{ab},1} \\ V^{\mathsf{a}}_{\mathsf{ac},1} \\ \mathbf{0}_{1\times 4} \\ \mathbf{0}_{2\times 4} \\ V^{\mathsf{a}}_{\mathsf{ab},2} \\ V^{\mathsf{a}}_{\mathsf{ac},2} \\ \mathbf{0}_{1\times 4} \\ \mathbf{0}_{2\times 4} \end{bmatrix} \underbrace{\begin{bmatrix} \mathsf{A}_1 \\ \mathsf{A}_2 \\ \mathsf{A}_3 \\ \mathsf{A}_4 \end{bmatrix}}_{\mathsf{A}_{[4]}} + \begin{bmatrix} V^{\mathsf{b}}_{\mathsf{ab},1} \\ \mathbf{0}_{1\times 4} \\ V^{\mathsf{b}}_{\mathsf{bc},1} \\ \mathbf{0}_{2\times 4} \\ V^{\mathsf{b}}_{\mathsf{ab},2} \\ \mathbf{0}_{1\times 4} \\ V^{\mathsf{b}}_{\mathsf{bc},2} \\ \mathbf{0}_{2\times 4} \end{bmatrix} \underbrace{\begin{bmatrix} \mathsf{B}_1 \\ \mathsf{B}_2 \\ \mathsf{B}_3 \\ \mathsf{B}_4 \end{bmatrix}}_{\mathsf{B}_{[4]}} + \begin{bmatrix} \mathbf{0}_{1\times 4} \\ V^{\mathsf{c}}_{\mathsf{ac},1} \\ V^{\mathsf{c}}_{\mathsf{bc},1} \\ \mathbf{0}_{2\times 4} \\ \mathbf{0}_{1\times 4} \\ V^{\mathsf{c}}_{\mathsf{ac},2} \\ V^{\mathsf{c}}_{\mathsf{bc},2} \\ \mathbf{0}_{2\times 4} \end{bmatrix} \underbrace{\begin{bmatrix} \mathsf{C}_1 \\ \mathsf{C}_2 \\ \mathsf{C}_3 \\ \mathsf{C}_4 \end{bmatrix}}_{\mathsf{C}_{[4]}} + \begin{bmatrix} \mathbf{0}_{3\times 4} \\ V^{\mathsf{d}}_{\mathsf{d},1} \\ V^{\mathsf{d}}_{\mathsf{d},2} \\ \mathbf{0}_{3\times 4} \\ V^{\mathsf{d}}_{\mathsf{d},3} \\ V^{\mathsf{d}}_{\mathsf{d},4} \end{bmatrix} \underbrace{\begin{bmatrix} \mathsf{D}_1 \\ \mathsf{D}_2 \\ \mathsf{D}_3 \\ \mathsf{D}_4 \end{bmatrix}}_{\mathsf{D}_{[4]}} \tag{25}$$

which indicates the precoding operations at each server with an encoding matrix V_{-}^{-}. For example, Server $\mathcal{S}_{\mathsf{ab}}$, precodes the $L \times 1 = 4 \times 1$ vector of data stream A (denoted as $\mathsf{A}_{[4]}$) with the $2N_{\mathsf{ab}} \times L = 2 \times 4$ precoding matrix $V^{\mathsf{a}}_{\mathsf{ab}}$, whose i^{th} row is denoted by $V^{\mathsf{a}}_{\mathsf{ab},i}$. Similarly, Server $\mathcal{S}_{\mathsf{ab}}$ precodes data stream B with the 2×4 precoding matrix $V^{\mathsf{b}}_{\mathsf{ab}}$. The precoded symbols are then mapped to the inputs controlled by Server ab, i.e., x_1, x_6, so that we have,

$$\begin{bmatrix} x_1 \\ x_6 \end{bmatrix} = V^{\mathsf{a}}_{\mathsf{ab}}\mathsf{A}_{[4]} + V^{\mathsf{b}}_{\mathsf{ab}}\mathsf{B}_{[4]}. \tag{26}$$

Each server similarly precodes the data streams available to it with its corresponding precoding matrices. Fig. 2 illustrates the precoding and decoding operations.

The precoding matrices are now specified as,

$$\begin{aligned} &\begin{bmatrix} V^{\mathsf{a}}_{\mathsf{ab}} \\ V^{\mathsf{a}}_{\mathsf{ac}} \end{bmatrix} = \left(V_{\text{dec}}\mathbf{M}_{(1,6,2,7)}\right)^{-1}, \ \begin{bmatrix} V^{\mathsf{b}}_{\mathsf{ab}} \\ V^{\mathsf{b}}_{\mathsf{bc}} \end{bmatrix} = \left(V_{\text{dec}}\mathbf{M}_{(1,6,3,8)}\right)^{-1}, \\ &\begin{bmatrix} V^{\mathsf{c}}_{\mathsf{ac}} \\ V^{\mathsf{c}}_{\mathsf{bc}} \end{bmatrix} = \left(V_{\text{dec}}\mathbf{M}_{(2,7,3,8)}\right)^{-1}, \ V^{\mathsf{d}}_{\mathsf{d}} = \left(V_{\text{dec}}\mathbf{M}_{(4,5,9,10)}\right)^{-1}. \end{aligned} \tag{27}$$

where $\mathbf{M}_{(i_1,i_2,\cdots,i_n)}$ is an $N \times n$ submatrix of $\mathbf{M}$ comprised of the $(i_1, i_2, \cdots, i_n)^{th}$ columns of $\mathbf{M}$. It is easy to verify that $\det(V_{\text{dec}}\mathbf{M}_{(1,6,2,7)}) = \det(V_{\text{dec}}\mathbf{M}_{(2,7,3,8)}) = 1$ and $\det(V_{\text{dec}}\mathbf{M}_{(1,6,3,8)}) = \det(V_{\text{dec}}\mathbf{M}_{(4,5,9,10)}) = -1$, thus all 4 inverses in (27) exist. With all choices explicitly specified, it is similarly easy to verify that we have,

$$V_{\text{dec}}\mathbf{y} = V_{\text{dec}}\mathbf{M}\mathbf{x} = \mathsf{A}_{[4]} + \mathsf{B}_{[4]} + \mathsf{C}_{[4]} + \mathsf{D}_{[4]}. \tag{28}$$

Thus, Alice is able to compute 4 instances of the desired sum, with the total download cost of 5 qudits. The coding scheme achieves the rate $4/5$ qudits/computation, matching the capacity of this Σ-QMAC setting.

[8] In fact, any $\mathbf{M} \in \mathbb{F}_d^{N\times 2N}$ of the form $[\mathbf{I}_N, \mathbf{S}_N]$ where $\mathbf{S}_N = \mathbf{S}_N^T$ satisfies $\mathbf{M}\mathbf{J}_{2N}\mathbf{M} = \mathbf{0}_{N\times N}$ and is therefore a valid N-sum box [36].

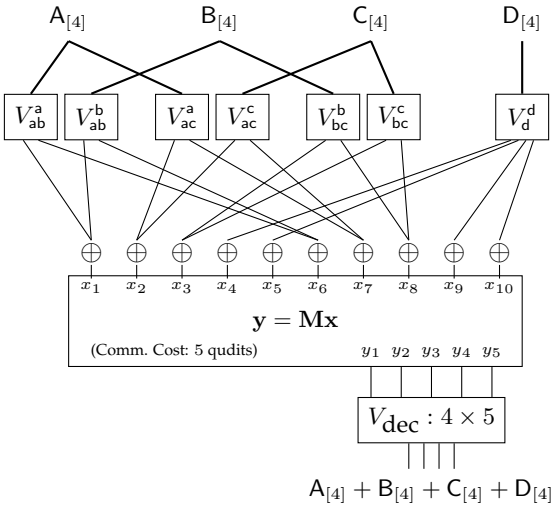

Figure 2: Precoding and decoding structure to achieve fully-entangled capacity C^{fullent} for the data replication map of Fig. 1.

6.2 Example: Symmetric Σ-QMAC with $S = 8$

Setting $S = 8$, for $\alpha, \beta \in \{1, 2, \cdots, 8\}$, we show the values of $C_\alpha^{(\beta)}$ in the following table according to Corollary 4. The first column ($\beta = 1$) corresponds to the fully-unentangled capacities. The

Table 2: $C_\alpha^{(\beta)}$ for $S = 8$.

$C_\alpha^{(\beta)}$ α \ β	1	2	3	4	5	6	7	8
1	1/8	1/4	1/4	1/4	1/4	1/4	1/4	1/4
2	1/4	13/28	13/28	1/2	1/2	1/2	1/2	1/2
3	3/8	9/14	9/14	5/7	5/7	3/4	3/4	3/4
4	1/2	11/14	11/14	61/70	61/70	13/14	13/14	1
5	5/8	25/28	25/28	27/28	27/28	1	1	1
6	3/4	27/28	27/28	1	1	1	1	1
7	7/8	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1

second column ($\beta = 2$) corresponds to the 2-party-entangled capacities. The capacities in this column are achievable by the 2-sum protocol. The last column ($\beta = 8$) corresponds to the fully-entangled capacities. Comparing the columns corresponding to $\beta = 1$ and $\beta = 8$, note that the DSC gain is 2 provided that the capacity does not exceed 1 dit/qudit. Also, note that the bipartite entanglement (i.e., $\beta = 2$) is in general not enough to achieve the maximal DSC gain.

6.3 Example: Minimizing the Maximal Entanglement

Maintaining entanglement across many parties tends to be increasingly challenging, as more parties become involved. So it is desirable to have smaller cliques without losing the capacity, moti-

vating the problem of identifying entanglement distribution maps $\mathcal{E}$, with the maximal clique size as small as possible such that the capacity achieved with $\mathcal{E}$ is the same as C^{fullent}. Mathematically, the problem is to find $\beta^* \triangleq \min\{\beta \mid C^{(\beta)} = C^{\text{fullent}}\}$ for a fixed data replication map. For example, consider the data distribution $\mathcal{W}$ shown in Fig. 1. We can see from Table 1 that the smallest value of β for this example is 4, same as S, i.e., all servers need to be entangled, because even if every subset of 3 of the 4 servers has an entangled system, the capacity is still only $3/4$, which is still less than $C^{\text{fullent}} = 4/5$. However, as evident from Table 2, where the data replication map is symmetric, it is in general not necessary to have all servers entangled in order to achieve the fully-entangled capacity C^{fullent}. For example, for $S = 8$ servers and $\alpha = 3$, we have $C_3^{(6)} = C_3^* = 3/4$, which means that it suffices to have $\beta = 6$ in order to achieve the fully-entangled capacity. Therefore, based on Table 2, for the cases with symmetric data replication, i.e., $K = \binom{S}{\alpha}$ and $\mathcal{W} : [K] \leftrightarrow \binom{[S]}{\alpha}$, we have $\beta^* = 2, 4, 6, 8, 6, 4, 2, 1$ for $\alpha = 1, 2, 3, 4, 5, 6, 7, 8$. From Corollary 4 it can further be verified that for general S,

$$\beta^* = \begin{cases} 2\alpha, & \alpha \leq \lfloor S/2 \rfloor, \\ 2(S-\alpha), & \lceil S/2 \rceil \leq \alpha \leq S-1, \\ 1, & \alpha = S. \end{cases} \tag{29}$$

The proof of (29) can be found in Appendix C. The intuition that emerges from this is that both extremes of too much data replication (large α) and too little data replication (small α) require relatively little entanglement (small β^*) to achieve their maximal DSC gain, rather the intermediate regimes of data replication are the ones that require the most entanglement to maximize their DSC gain.

6.4 Example: 2-sum Protocol Based Coding for Fig. 1

The main purpose of this example is to illustrate Corollary 5. Note that there are two Σ-QMAC problems involved in Corollary 5. In this example, let us again consider the data replication map defined in Fig. 1. We consider the following two Σ-QMAC problems.

1. $\mathcal{P}$: The Σ-QMAC problem with data replication map $\mathcal{W}$ as shown in Fig. 1. Specifically, there are four data streams, denoted as $\mathsf{A}, \mathsf{B}, \mathsf{C}$ and D. The four servers are denoted as $\mathcal{S}_{\mathsf{ab}}, \mathcal{S}_{\mathsf{ac}}, \mathcal{S}_{\mathsf{bc}}$ and $\mathcal{S}_{\mathsf{d}}$. We wish to find the 2-party-entangled capacity $C^{(2)}(\mathcal{W})$.

2. $\widetilde{\mathcal{P}}$: The (hypothetical) Σ-QMAC problem with data replication map $\widetilde{\mathcal{W}}$. Specifically, there are $\binom{4}{2} = 6$ servers, each has the data streams that are available to a unique pair of servers in $\mathcal{P}$. Therefore, the data streams available to the servers in $\widetilde{\mathcal{P}}$ are $\mathsf{ABC}, \mathsf{ABC}, \mathsf{ABD}, \mathsf{ABC}, \mathsf{ACD}, \mathsf{BCD}$, respectively. Without loss of generality, the three servers that know ABC can be considered as one server. We refer to the server that has data streams ABC as $\mathcal{S}_{\mathsf{abc}}$, and similarly we define the rest 3 servers as $\mathcal{S}_{\mathsf{abd}}, \mathcal{S}_{\mathsf{acd}}$ and $\mathcal{S}_{\mathsf{bcd}}$. We wish to find the fully-unentangled capacity $C^{\text{unent}}(\widetilde{\mathcal{W}})$.

According to Corollary 5, these two Σ-QMAC problems have the same capacity, i.e., $C^{(2)}(\mathcal{W}) = C^{\text{unent}}(\widetilde{\mathcal{W}})$, both equal to $3/4$ by Corollary 2. A scheme for the problem $\widetilde{\mathcal{P}}$ is illustrated in Table 3. Note that in $\widetilde{\mathcal{P}}$, no entanglement across different servers is allowed, and our scheme in Table 3 simply treats qudits as dits. Since $Y_{\mathsf{abc}} + Y_{\mathsf{abd}} = \mathsf{A}_1 + \mathsf{B}_1 + \mathsf{C}_1 + \mathsf{D}_1$, $Y_{\mathsf{abd}} + Y_{\mathsf{acd}} = \mathsf{A}_2 + \mathsf{B}_2 + \mathsf{C}_2 + \mathsf{D}_2$ and $Y_{\mathsf{acd}} + Y_{\mathsf{bcd}} = \mathsf{A}_3 + \mathsf{B}_3 + \mathsf{C}_3 + \mathsf{D}_3$, the scheme allows Alice to compute $L = 3$ instances of

Server	Transmission
$\mathcal{S}_{\mathsf{abc}}$	$Y_{\mathsf{abc}} = (\mathsf{A}_1 - \mathsf{A}_2 + \mathsf{A}_3) + (\mathsf{B}_1 - \mathsf{B}_2) + \mathsf{C}_1$
$\mathcal{S}_{\mathsf{abd}}$	$Y_{\mathsf{abd}} = (\mathsf{A}_2 - \mathsf{A}_3) + \mathsf{B}_2 + \mathsf{D}_1$
$\mathcal{S}_{\mathsf{acd}}$	$Y_{\mathsf{acd}} = \mathsf{A}_3 + \mathsf{C}_2 + (\mathsf{D}_2 - \mathsf{D}_1)$
$\mathcal{S}_{\mathsf{bcd}}$	$Y_{\mathsf{bcd}} = \mathsf{B}_3 + (\mathsf{C}_3 - \mathsf{C}_2) + (\mathsf{D}_3 - \mathsf{D}_2 + \mathsf{D}_1)$

Table 3: Coding scheme for $\widetilde{\mathcal{P}}$

the desired sum with a total cost of 4 qudits (Y_{abc}, Y_{abd}, Y_{acd} and Y_{bcd} each costs one qudit). The capacity $3/4$ is thus achieved.

Fig. 3 shows that a scheme thats achieve $C^{(2)}(\mathcal{W}) = 3/4$ in the problem $\mathcal{P}$ can be deduced from Table 3, which allows Alice to compute 6 instances of the sum in the problem $\mathcal{P}$, with 4 uses of the 2-sum protocols, thus achieving the rate $6/8 = 3/4$. Note that each use of the 2-sum

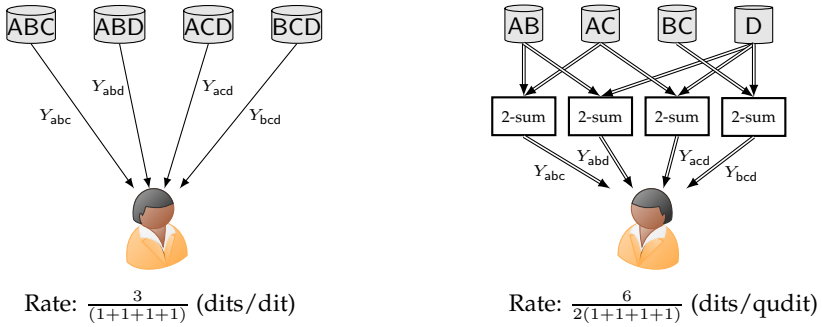

Figure 3: A comparison of the schemes that achieve $C^{\text{unent}}(\widetilde{\mathcal{W}}) = 3/4$ in the problem $\widetilde{\mathcal{P}}$ (LHS) and $C^{(2)}(\mathcal{W}) = 3/4$ in the problem $\mathcal{P}$ (RHS). In RHS, each of Y_{abc}, Y_{abd}, Y_{acd} and Y_{bcd} contains two instances, e.g., $Y_{\mathsf{abc}} = (Y^{(1)}_{\mathsf{abc}}, Y^{(2)}_{\mathsf{abc}})$, where $Y^{(1)}_{\mathsf{abc}}$ is the function of $(\mathsf{A}_i, \mathsf{B}_i, \mathsf{C}_i, \mathsf{D}_i)_{i=1}^{3}$ as shown in Table 3 and $Y^{(2)}_{\mathsf{abc}}$ is the corresponding function of $(\mathsf{A}_i, \mathsf{B}_i, \mathsf{C}_i, \mathsf{D}_i)_{i=4}^{6}$.

protocol transmits 2 instances of the symbol that is sent by a server in the problem $\widetilde{\mathcal{P}}$. Specifically, for example, the two servers $\mathcal{S}_{\mathsf{ab}}$ and $\mathcal{S}_{\mathsf{ac}}$ in the problem $\mathcal{P}$ use the 2-sum protocol once, with the two inputs at $\mathcal{S}_{\mathsf{ab}}$ specified as $(\mathsf{A}_1 - \mathsf{A}_2 + \mathsf{A}_3)$ and $(\mathsf{A}_4 - \mathsf{A}_5 + \mathsf{A}_6)$, the two inputs at $\mathcal{S}_{\mathsf{ac}}$ specified as $(\mathsf{B}_1 - \mathsf{B}_2) + \mathsf{C}_1$ and $(\mathsf{B}_4 - \mathsf{B}_5) + \mathsf{C}_4$, so that Alice gets $Y^{(1)}_{\mathsf{abc}} = (\mathsf{A}_1 - \mathsf{A}_2 + \mathsf{A}_3) + (\mathsf{B}_1 - \mathsf{B}_2) + \mathsf{C}_1$ and $Y^{(2)}_{\mathsf{abc}} = (\mathsf{A}_4 - \mathsf{A}_5 + \mathsf{A}_6) + (\mathsf{B}_4 - \mathsf{B}_5) + \mathsf{C}_4$.

6.5 3-party entanglement is not necessary in Σ-QMAC

As an example to illustrate Corollary 7, let us once again consider the data replication map in Fig. 1. One can quickly check from Table 1 that for any capacity that is associated with an entanglement distribution map that contains a 3-party clique, the same capacity is achievable for another entanglement distribution map where the 3-party clique is replaced with three 2-party cliques, each containing a unique pair of servers in the 3-party cliques. Another example is the symmetric settings with $S = 8$ servers as shown in Table 2. Note that in each row, the 3-party-entangled capacity is always equal to the 2-party-entangled capacity, meaning that 3-party entanglement is

not necessary for achieving a higher capacity.

6.6 The necessity of S-party entanglement for Σ-QMAC ($S \neq 3$)

The symmetric setting in Section 6.2 may support the intuition that β-partite entanglement is unnecessary for odd β, i.e., $C^{(\beta)}(\mathcal{W}) = C^{(\beta-1)}(\mathcal{W})$, because the columns corresponding to odd values of β in Table 2 are identical to their preceding columns. The intuition may even be strengthened by Corollary 7 which shows that indeed $C^{(3)}(\mathcal{W}) = C^{(2)}(\mathcal{W})$ for any $\mathcal{W}$. Perhaps surprisingly then, Corollary 8 reveals that $\beta = 3$ is only an exception and it is not generally true that $C^{(\beta)}(\mathcal{W}) = C^{(\beta-1)}(\mathcal{W})$ for all data replication patterns $\mathcal{W}$ if β is an odd number.

Leaving the proof of Corollary 8 to Appendix F, let us consider here the case $\beta = S = 5$ to see that indeed there exists a data replication pattern $\mathcal{W}$ such that $C^{(5)}(\mathcal{W}) > C^{(4)}(\mathcal{W})$. Let $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ and E denote $K = 5$ data streams. The data replication map $\mathcal{W}$ is such that Server $\mathcal{S}_1$ has $\mathsf{A}, \mathsf{B}, \mathsf{C}$, Server $\mathcal{S}_2$ has $\mathsf{A}, \mathsf{B}, \mathsf{D}$, Server $\mathcal{S}_3$ has $\mathsf{A}, \mathsf{C}, \mathsf{D}$, Server $\mathcal{S}_4$ has $\mathsf{B}, \mathsf{C}, \mathsf{D}$ and Server $\mathcal{S}_5$ has E. If we only look at the first 4 servers, this is the symmetric data replication map with 4 data streams, each being replicated in a unique subset of 3 servers. The asymmetry comes from the additional data stream, E, that is only available at server $\mathcal{S}_5$. It can be verified by Theorem 1 that for this data replication map, $C^{\text{fullent}}(\mathcal{W}) \triangleq C^{(5)}(\mathcal{W}) = 6/7$ and $C^{(4)}(\mathcal{W}) = 5/6$, which together show that the 5-partite entanglement is necessary for this $\mathcal{W}$.

6.7 Separability of Σ-QMAC

Let us provide some examples to illustrate the notion of separability for Corollary 9. Define data replication maps $\mathcal{W}_1, \mathcal{W}_2$ and $\mathcal{W}_3$ such that

$$\begin{aligned}
&\mathcal{W}_1(1) = \{1,2\},\ \mathcal{W}_1(2) = \{1,3\},\ \mathcal{W}_1(3) = \{2,3\}\\
&\mathcal{W}_2(4) = \{4\}\\
&\mathcal{W}_3(k) = \begin{cases} \mathcal{W}_1(k), & k \in \{1,2,3\} \\ \mathcal{W}_2(k), & k = 4 \end{cases}
\end{aligned}$$

For $i \in \{1,2,3\}$, let Σ-QMAC$_i$ be the Σ-QMAC problem with data replication map $\mathcal{W}_i$ and fully-entangled capacity $C^{\text{fullent}}(\mathcal{W}_i)$. Note that $\mathcal{W}_3$ corresponds to the data replication map shown in Fig. 1. By Corollary 1, we have $1/C^{\text{fullent}}(\mathcal{W}_1) = 1$, $1/C^{\text{fullent}}(\mathcal{W}_2) = 1$ and $1/C^{\text{fullent}}(\mathcal{W}_3) = 5/4$. This shows that Σ-QMAC$_1$ and Σ-QMAC$_2$ are not separable.

Next let us provide an example where two Σ-QMACs are separable. Define data replication maps $\mathcal{W}'_1, \mathcal{W}'_2$ and $\mathcal{W}'_3$ such that

$$\begin{aligned}
&\mathcal{W}'_1(1) = \{1\},\ \mathcal{W}'_1(2) = \{2\}\\
&\mathcal{W}'_2(3) = \{3\},\ \mathcal{W}'_2(4) = \{4\}\\
&\mathcal{W}'_3(k) = \begin{cases} \mathcal{W}'_1(k), & k \in \{1,2\} \\ \mathcal{W}'_2(k), & k \in \{3,4\} \end{cases}
\end{aligned}$$

For $i \in \{1,2,3\}$, let Σ-QMAC$'_i$ be the Σ-QMAC problem with data replication map $\mathcal{W}'_i$ and fully-entangled capacity $C^{\text{fullent}}(\mathcal{W}'_i)$. By Corollary 1, we have $1/C^{\text{fullent}}(\mathcal{W}'_1) = 1$, $1/C^{\text{fullent}}(\mathcal{W}'_2) = 1$ and $1/C^{\text{fullent}}(\mathcal{W}'_3) = 2$. This shows that Σ-QMAC$'_1$ and Σ-QMAC$'_2$ are separable.

7 Proof of Theorem 1: Converse

Our converse bound for proving Theorem 1 is based on the cut-set argument with the capacity result of classical-quantum communication channel (e.g., [18, 59]). The definitions of quantum coding schemes, feasible region, capacity and the DSC gain follow from those of the Σ-QMAC.

7.1 Prerequisite: Dense Coding Capacity

Consider a point to point quantum communication setting with a sender, Bob, and a receiver, Alice. Quantum systems $\mathcal{Q}_A$ and $\mathcal{Q}_B$ are provided to Alice and Bob, respectively. We use $|\mathcal{Q}|$ to denote the dimension of a quantum system $\mathcal{Q}$. The composite system $\mathcal{Q}_B\mathcal{Q}_A$ is in the initial state ρ^{BA}, described by the density operator. Independent of ρ^{BA} there is a random variable X, so that with probability $p_X(x)$, Bob applies a unitary operation U_x on $\mathcal{Q}_B$. The resulting state of the system $\mathcal{Q}_B\mathcal{Q}_A$ is thus $\rho_x^{BA'} = (U_B \otimes I_A)\rho^{BA}(U_B^\dagger \otimes I_A)$ for $X = x$. Then Bob sends $\mathcal{Q}_B$ to Alice, which allows Alice to measure $\mathcal{Q}_B\mathcal{Q}_A$ and obtain the outcome Y. The *dense coding capacity* [18], defined as the maximum amount of information that Alice can learn about X from Y, is equal to $\max I(X;Y)$ where the maximum is taken over all p_X and all possible measurement at Alice. This value can be strictly larger than $\log_d|\mathcal{Q}_B|$ (dits), in which case the coding scheme is called a *dense* coding. As shown by [18], $\max I(X;Y) = \log_d|\mathcal{Q}_B| + S(\rho^A) - S(\rho^{BA})$ (dits), where $S(\cdot)$ denotes the von Neumann entropy and ρ^A is the reduced density operator for $\mathcal{Q}_A$. Due to the inequalities $|S(\rho^A) - S(\rho^B)| \le S(\rho^{BA})$, $S(\mathcal{Q}_A) \le \log_d|\mathcal{Q}_A|$ and non-negativity of von Neumann entropy, we obtain that

$$I(X;Y) \le \min(\log_d|\mathcal{Q}_B\mathcal{Q}_A|, 2\log_d|\mathcal{Q}_B|) \text{ dits.} \tag{30}$$

7.2 Proof of Converse

Consider any feasible coding scheme specified by

$$\big(L,\ ((\delta_{t,s})_{s\in\mathcal{E}(t)})_{t\in[T]},\ \rho_{[T]},\ ((\Phi_{t,s})_{s\in\mathcal{E}(t)})_{t\in[T]},\ (\{M_{t,y}\}_{y\in\mathcal{Y}_t})_{t\in[T]},\ \Psi\big). \tag{31}$$

Lemma 1 (Conditional Independence)**.** *Given any feasible scheme,* $\mathsf{Y}_1, \mathsf{Y}_2, \cdots, \mathsf{Y}_T$ *are mutually independent conditioned on the event* $(\mathsf{W}_1^{[L]} = w_1, \mathsf{W}_2^{[L]} = w_2, \cdots, \mathsf{W}_K^{[L]} = w_K)$ *for any* $w_1, w_2, \cdots, w_K \in \mathbb{F}_d^L$.

Proof. Recall that Alice receives the quantum system $\mathcal{Q}_t$ in the state $\rho_t' = (\otimes_{s\in\mathcal{E}(t)}U_{t,s})\rho_t(\otimes_{s\in\mathcal{E}(t)}U_{t,s})^\dagger$ for $t \in [T]$. Thus, for any $w_1, w_2, \cdots, w_K \in \mathbb{F}_d^L$,

$$\mathsf{Pr}(\mathsf{Y}_t = y_t \mid \mathsf{W}_1^{[L]} = w_1, \cdots, \mathsf{W}_K^{[L]} = w_K) = \mathsf{tr}(\rho_t' M_{t,y_t}),\ \ \forall t \in [T] \tag{32}$$

and

$$\begin{aligned}
&\mathsf{Pr}\big(\mathsf{Y}_1 = y_1, \cdots, \mathsf{Y}_T = y_T \mid \mathsf{W}_1^{[L]} = w_1, \cdots, \mathsf{W}_K^{[L]} = w_K\big) \\
&= \mathsf{tr}\big((\rho_1' \otimes \cdots \otimes \rho_T')(M_{1,y_1} \otimes \cdots \otimes M_{T,y_T})\big) &(33)\\
&= \mathsf{tr}\big((\rho_1' M_{1,y_1}) \otimes \cdots \otimes (\rho_T' M_{T,y_T})\big) &(34)\\
&= \prod_{t=1}^{T} \mathsf{tr}(\rho_t' M_{t,y_t}) &(35)
\end{aligned}$$

$$= \prod_{t=1}^{T} \mathsf{Pr}(Y_t = y_t \mid \mathsf{W}_1^{[L]} = w_1, \cdots, \mathsf{W}_K^{[L]} = w_K) \tag{36}$$

where the last step uses (32). It follows that $\mathsf{Y}_1, \mathsf{Y}_2, \cdots, \mathsf{Y}_T$ are conditionally independent. □

Let $\big(\mathsf{W}_1^{(\ell)}, \mathsf{W}_2^{(\ell)}, \cdots, \mathsf{W}_K^{(\ell)}\big)$ be the ℓ^{th} instance of the data streams. Since any feasible scheme must guarantee successful decoding for all realizations of $\big(\mathsf{W}_1^{(\ell)}, \mathsf{W}_2^{(\ell)}, \cdots, \mathsf{W}_K^{(\ell)}\big) \in \mathbb{F}_d^K$ for all ℓ, it must still guarantee successful decoding if we assume $\big(\mathsf{W}_1^{(\ell)}, \mathsf{W}_2^{(\ell)}, \cdots, \mathsf{W}_K^{(\ell)}\big)$ to be uniform over $\mathbb{F}_d^K$ for any $\ell \in [L]$, and independent over $\ell \in [L]$. For $t \in [T]$, let us account for the separate measurement corresponding to clique $\mathcal{E}(t)$. Following a regular cut set argument, for $k \in [K]$, denote $\mathcal{A}_{t,k} \triangleq \big([S] \backslash \mathcal{W}(k)\big) \cap \mathcal{E}(t)$ and let Servers $s \in \mathcal{A}_{t,k}$ join Alice as the receiver by bringing their quantum resource and data. Let $\mathcal{B}_{t,k} \triangleq \mathcal{E}(t) \backslash \mathcal{A}_{t,k} = \mathcal{E}(t) \cap \mathcal{W}(k)$ and consider Servers $s \in \mathcal{B}_{t,k}$ collectively as the transmitter. Denote the subsystem of $\mathcal{Q}_t$ that is sent from Servers $s \in \mathcal{B}_{t,k}$ as $\mathcal{Q}_B$ and the quantum subsystem of $\mathcal{Q}_t$ that is brought from Servers $s \in \mathcal{A}_{t,k}$ as $\mathcal{Q}_A$. Since the K data streams are mutually independent, conditioned on $\mathsf{W}_{[K]\backslash\{k\}}^{[L]}$, it follows from (30) that,

$$\begin{aligned}
&I\big(\mathsf{W}_k^{[L]}; \mathsf{Y}_t \,\big|\, \mathsf{W}_{[K]\backslash\{k\}}^{[L]}\big) \\
&\leq \min\big(\log_d |\mathcal{Q}_B \mathcal{Q}_A|, 2\log_d |\mathcal{Q}_B|\big) && (37)\\
&= \min\Big(\log_d \prod_{s\in\mathcal{E}(t)} |\mathcal{Q}_{t,s}|, 2\log_d \prod_{s\in\mathcal{B}_{t,k}} |\mathcal{Q}_{t,s}|\Big) && (38)\\
&= \min\Big(\sum_{s\in\mathcal{E}(t)} \log_d |\mathcal{Q}_{t,s}|, 2\sum_{s\in\mathcal{B}_{t,k}} |\mathcal{Q}_{t,s}|\Big) && (39)
\end{aligned}$$

where Y_t denotes the result after measuring the quantum system $\mathcal{Q}_B\mathcal{Q}_A$, i.e., the composite system comprised of $\mathcal{Q}_B$ and $\mathcal{Q}_A$.

We then have

$$\begin{aligned}
&\sum_{t\in[T]} I\big(\mathsf{W}_k^{[L]}; \mathsf{Y}_t \,\big|\, \mathsf{W}_{[K]\backslash\{k\}}^{[L]}\big) \\
&= \sum_{t\in[T]} H\big(\mathsf{Y}_t \,\big|\, \mathsf{W}_{[K]\backslash\{k\}}^{[L]}\big) - \sum_{t\in[T]} H\big(\mathsf{Y}_t \,\big|\, \mathsf{W}_{[K]}^{[L]}\big) && (40)\\
&\geq H\big(\mathsf{Y}_{[T]} \,\big|\, \mathsf{W}_{[K]\backslash\{k\}}^{[L]}\big) - \sum_{t\in[T]} H\big(\mathsf{Y}_t \,\big|\, \mathsf{W}_{[K]}^{[L]}\big) && (41)\\
&= H\big(\mathsf{Y}_{[T]} \,\big|\, \mathsf{W}_{[K]\backslash\{k\}}^{[L]}\big) - \sum_{t\in[T]} H\big(\mathsf{Y}_t \,\big|\, \mathsf{W}_{[K]}^{[L]}, \mathsf{Y}_{[t-1]}\big) && (42)\\
&= H\big(\mathsf{Y}_{[T]} \,\big|\, \mathsf{W}_{[K]\backslash\{k\}}^{[L]}\big) - H\big(\mathsf{Y}_{[T]} \,\big|\, \mathsf{W}_{[K]}^{[L]}\big) && (43)\\
&= I\big(\mathsf{W}_k^{[L]}; \mathsf{Y}_{[T]} \,\big|\, \mathsf{W}_{[K]\backslash\{k\}}^{[L]}\big) && (44)\\
&= H\big(\mathsf{W}_k^{[L]} \,\big|\, \mathsf{W}_{[K]\backslash\{k\}}^{[L]}\big) && (45)\\
&= H\big(\mathsf{W}_k^{[L]}\big) && (46)\\
&= L \text{ (dits)} && (47)
\end{aligned}$$

where Step (42) makes use of the conditional independence of $\mathsf{Y}_1, \mathsf{Y}_2, \cdots, \mathsf{Y}_T$ as implied by Lemma 1. Step (45) holds because conditioned on $\mathsf{W}_{[K]\setminus\{k\}}^{[L]}$, Alice must be able to recover $\mathsf{W}_k^{[L]}$ from $\mathsf{Y}_{[T]}$.

Combining (39) and (47), we obtain that for any $k \in [K]$,

$$\sum_{t\in[T]} \min\Big(\sum_{s\in\mathcal{E}(t)} \log_d |\mathcal{Q}_{t,s}|, 2 \sum_{s\in\mathcal{B}_{t,k}} |\mathcal{Q}_{t,s}| \Big) \geq L \text{ (dits)}. \tag{48}$$

Dividing by L on both sides gives us

$$\sum_{t\in[T]} \min\left\{ \sum_{s\in\mathcal{E}(t)} \Delta_{t,s}, \sum_{s\in\mathcal{B}_{t,k}} 2\Delta_{t,s} \right\} \geq 1, \ \forall k \in [K], \tag{49}$$

which matches the condition of the feasible region in Theorem 1.

8 Proof of Theorem 1: Achievability

8.1 Prerequisite: The N-sum Box

Building on the stabilizer formalism and quantum error correction literature on stabilizer codes, an implicit generalization of the 2-sum protocol is presented in [38], and subsequently crystallized as an N-sum box abstraction in [36]. The N-sum box has $2N$ classical inputs, labeled $x_1, x_2, \cdots, x_{2N} \in \mathbb{F}_d$, and N classical outputs $y_1, y_2, \cdots, y_N \in \mathbb{F}_d$, related by a MIMO MAC channel formulation as,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} M_{1,1} & \cdots & M_{1,2N} \\ \vdots & \vdots & \vdots \\ M_{N,1} & \cdots & M_{N,2N} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{2N} \end{bmatrix} \tag{50}$$

which can be represented compactly as $\mathbf{y} = \mathbf{M}\mathbf{x}$. The N-sum box abstraction represents the setting where N entangled qudits are distributed among K transmitters, such that each transmitter can perform conditional quantum X, Z gate operations on its qudit(s) to encode classical information. The transmitter that has the n^{th} qudit controls the inputs x_n and x_{N+n} of the N-sum box. For example, if Qudits 1 and 3 are given to Transmitter 1, then in the N-sum box abstraction the inputs $x_1, x_{1+N}, x_3, x_{3+N}$ are the inputs available to Transmitter 1. The N outputs are the result of the quantum measurement performed by Alice. Since the N qudits are sent to Alice for the quantum measurement, the N-sum box has a quantum communication cost of N qudits. Now let us consider the channel matrix $\mathbf{M}$. Different choices of entanglement states and quantum-measurement bases produce different channel matrices. Depending on the desired computation task a suitable $\mathbf{M}$ may be chosen from the set of feasible choices. The channel matrices that can be obtained from the stabilizer-based construction are precisely those (see [36]) that are strongly self-orthogonal, i.e., that satisfy the following two conditions,

$$\text{rk}(\mathbf{M}) = N, \qquad\qquad \mathbf{M}\mathbf{J}_{2N}\mathbf{M}^T = \mathbf{0}_{N\times N} \tag{51}$$

where $\mathbf{J}_{2N} = \begin{pmatrix} \mathbf{0} & -\mathbf{I}_N \\ \mathbf{I}_N & \mathbf{0} \end{pmatrix}$ and $\mathbf{I}_N$ is the $N \times N$ identity matrix. Designing quantum-codes for the Σ-QMAC using the N-sum box abstraction entails a choice of not only which N-sum boxes to use,

how many of the inputs of each N-sum box to assign to each transmitter, and how to precode at each transmitter in the MIMO MAC for the desired computation, but in contrast to conventional (wireless) MIMO MAC settings where the channels are randomly chosen by nature, here we also have the freedom to design suitable channel matrices $\mathbf{M}$ for the desired computation task, within the class of feasible choices. The N-sum box abstraction then guarantees that corresponding to these choices there exist initial quantum entanglements, quantum-coding operations at the transmitters, and quantum-measurement operations at the receiver, that achieve the desired MIMO MAC functionality, at the communication cost of N qudits for each N-sum box utilized by the coding scheme.

8.2 Proof of Achievability

Our proof of achievability combines a series of results in network coding literature, together with the N-sum box formulation. Therefore, let us first summarize these results into the following lemmas to facilitate the proof. Consider the following setup. Let $K \in \mathbb{N}$. For $k \in [K]$, let $\mathbf{H}_k \in \mathbb{F}_q^{n\times m_k}$. Let W and $\mathsf{W}_k, k \in [K]$ be sources generating symbols in $\mathbb{F}_q$. Let us define the following two network coding type problems.

Sum-network: There is a MIMO multiple access channel with K transmitters and one receiver. The input at Transmitter k is $X_k \in \mathbb{F}_q^{m_k\times 1}$. The output at the receiver is $Y = \sum_{k\in[K]} \mathbf{H}_k X_k$. Transmitter k knows W_k. The receiver wants to know $\mathsf{W}_1 + \mathsf{W}_2 + \cdots + \mathsf{W}_K$. A feasible coding scheme can be described by $(L, N, \phi_{[K]}, \psi)$ so that the encoders ϕ_k map $\mathsf{W}_k^{[L]}$ to $X_k^{[N]}, \forall k \in [K]$, and the decoder ψ maps $Y^{[N]}$ to $\sum_{k\in[K]} \mathsf{W}_k^{[L]}$. A rate R is achievable if there exists a scheme so that $R \le L/N$.

Multicast: There is a MIMO broadcast channel with one transmitter and K receivers. The input at the Transmitter is $\widetilde{X} \in \mathbb{F}_q^{n\times 1}$. The output at Receiver k is $\widetilde{Y}_k = \mathbf{H}_k^T \widetilde{X}$. The transmitter knows a message W. All K receivers want to decode W. A feasible coding scheme can be described by $(L, N, \widetilde{\phi}, \widetilde{\psi}_{[K]})$ so that the encoder $\widetilde{\phi}$ maps $\mathsf{W}^{[L]}$ to $\widetilde{X}^{[N]}$, and the decoders $\widetilde{\psi}_k$ map $\widetilde{Y}_k^{[N]}$ to $\mathsf{W}^{[L]}, \forall k \in [K]$. A rate R is achievable if there exists a scheme so that $R \le L/N$.

Lemma 2 (Duality [28])**.** *If $(L, N, \phi_{[K]}, \psi)$ is a feasible linear coding scheme[9] for the sum-network, then there exists a feasible linear coding scheme $(L, N, \widetilde{\phi}, \widetilde{\psi}_{[K]})$ for the corresponding multicast problem, and vice versa.*

Lemma 3 (Multicast Capacity [60])**.** *The capacity (supreme of achievable rates) of the multicast problem is $\min_{k\in[K]} \mathrm{rk}(\mathbf{H}_k)$, and it can be achieved by linear coding schemes.*

We only need the achievability side of Lemma 3. Although the idea essentially follows from [60, 61], since our formulation here is slightly different, we provide an alternative proof of this lemma in Appendix G. Based on these two lemmas, the next corollary becomes obvious.

Corollary 10. *The rate $R = \min_{k\in[K]} \mathrm{rk}(\mathbf{H}_k)$ is achievable by a linear coding scheme in the sum-network.*

Going back to the proof, our achievable scheme is essentially based on the achievable scheme of a sum-network, which is constructed by the N-sum box formulation. Recall that there are in total T cliques (sets of servers that are allowed to share an entangled quantum system). For the t^{th} clique ($t \in [T]$), we let the servers $\mathcal{E}(t)$ implement an N_t-sum box in $\mathbb{F}_q = \mathbb{F}_{d^z}$, so that Server

[9] If the coding functions $\phi_{[K]}$ and decoding function ψ are linear functions, the scheme is said to be linear. The linearity for the Multicast setting is similarly defined.

$s \in \mathcal{E}(t)$ controls $2N_{t,s}$ inputs in $\mathbb{F}_q$, by using a quantum subsystem $\mathcal{Q}_{t,s}$ with dimension specified to $\delta_{t,s} = q^{N_{t,s}}$. Equivalently, $\mathcal{Q}_{t,s}$ can be considered as $N_{t,s}$ q-ary quantum subsystems, or qz qudits. $z \in \mathbb{N}$ is free to be chosen later.

For $t \in [T]$, the transfer matrix of the t^{th} box is denoted as $\mathbf{M}_t \in \mathbb{F}_q^{N_t \times 2N_t}$. Recall that for $k \in [K]$, each data stream W_k generates symbols in $\mathbb{F}_d$. Equivalently, we can consider these symbols as in $\mathbb{F}_q$ where $q = d^z$ for any $z \in \mathbb{N}$, by regarding each z symbols as one super-symbol in $\mathbb{F}_q$. Since we do not put any constraint on the batch size L, let $\mathbf{W}_k \in \mathbb{F}_d^{L' \times 1}$ denote the first L' symbols of W_k considered in $\mathbb{F}_q$ (which correspond to $L = L'z$ symbols in the original field $\mathbb{F}_d$, or $\mathsf{W}_k^{(L'z)}$).

Next let us define the input and the output of the t^{th} box. The input of the t^{th} box is $\mathbf{x}_t \in \mathbb{F}_q^{2N_t \times 1}$, and therefore the output of the t^{th} box is $\mathbf{y}_t = \mathbf{M}_t \mathbf{x}_t \in \mathbb{F}_q^{N_t \times 1}$. $\mathbf{x}_t$ consists of the $N_{t,s}$ pairs of inputs controlled by Servers $s \in \mathcal{E}(t)$. It is then obvious that the t^{th} box is a MIMO-MAC channel with all symbols defined in $\mathbb{F}_q$, and the input of each server $s \in \mathbb{E}(t)$ corresponds to some $2N_{t,s}$ columns of $\mathbf{M}_t$. We will also say that these $2N_{t,s}$ columns are controlled by Server s. A column of $\mathbf{M}_t$ is said to be *accessible* by a data stream $\mathsf{W}_k, k \in [K]$ if and only if this column is controlled by some server $s \in \mathcal{E}(t)$ that also knows this data stream, i.e., $s \in \mathcal{E}(t) \cap \mathcal{W}(k)$.

For $k \in [K]$, define $\mathbf{M}_{t,k}$ as the set of columns of $\mathbf{M}_t$ that are *accessible* by the k^{th} data stream. Note that $\mathbf{M}_{t,k}$ can be empty for some k, if the t^{th} clique does not contain any server that knows W_k. We claim that the output of the t^{th} box can be made as $\mathbf{y}_t = \sum_{k \in [K]} \mathbf{M}_{t,k} \phi_{t,k}(\mathbf{W}_k)$, where $\phi_{t,k}$ maps $\mathbf{W}_k$ to a vector in $\mathbb{F}_q$ with length $\sum_{s \in \mathcal{E}(t) \cap \mathcal{W}(k)} N_{t,s}$.

To prove the claim, for any $t \in [T], k \in [K]$, let $\phi_{t,s}^k : \mathbb{F}_q^{L' \times 1} \mapsto \mathbb{F}_q^{2N_{t,s} \times 1}$ describe a map if $s \in \mathcal{E}(t) \cap \mathcal{W}(k)$, and let the input for any one use of the t^{th} MIMO-MAC channel (i.e., the t^{th} box) at Server $s \in \mathcal{E}(t)$ be specified as $\sum_{k: s \in \mathcal{W}(k)} \phi_{t,s}^k(\mathbf{W}_k)$. The output of the t^{th} box is then determined as $\mathbf{y}_t \in \mathbb{F}_q^{N_t \times 1}$ such that,

$$\mathbf{y}_t = \mathbf{M}_t \mathbf{x}_t = \sum_{k \in [K]} \mathbf{M}_{t,k} \underbrace{\begin{bmatrix} \phi_{t,s_1}^k(\mathbf{W}_k) \\ \phi_{t,s_2}^k(\mathbf{W}_k) \\ \vdots \\ \phi_{t,s_n}^k(\mathbf{W}_k) \end{bmatrix}}_{\phi_{t,k}(\mathbf{W}_k)} \tag{52}$$

where $\{s_1, s_2, \cdots, s_n\} = \mathcal{E}(t) \cap \mathcal{W}(k)$, and $\mathbf{M}_{t,k}$ is a submatrix of $\mathbf{M}_t$ comprised of $\sum_{s \in \mathcal{E}(t) \cap \mathcal{W}(k)} 2N_{t,s}$ columns of $\mathbf{M}_t$ that are accessible by $\mathbf{W}_k$. Therefore, we obtain the general expression of the output $\mathbf{y}_t$ as in (52).

Note that we have in total T boxes and therefore T MIMO-MAC channels. We can equivalently consider the T channels as one big channel, and the output of the big channel (for one channel use) can be written as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{bmatrix} = \sum_{k \in [K]} \underbrace{\begin{bmatrix} \mathbf{M}_{1,k} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{2,k} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M}_{T,k} \end{bmatrix}}_{\overline{\mathbf{M}}_k} \underbrace{\begin{bmatrix} \phi_{1,k}(\mathbf{W}_k) \\ \phi_{2,k}(\mathbf{W}_k) \\ \vdots \\ \phi_{T,k}(\mathbf{W}_k) \end{bmatrix}}_{\phi_k(\mathbf{W}_k)}. \tag{53}$$

Note that there is no restriction on ϕ_k except the dimensions of its input and output. (53) thus describes the input-output relation of a K-transmitter MIMO MAC with the k^{th} transmitter knowing only W_k. According to Corollary 10, the receiver of this channel (Alice) can compute the sum

$\sum_{k\in[K]} \mathsf{W}_k$ (in $\mathbb{F}_q$) at the rate $\min_{k\in[K]} \mathrm{rk}_q(\overline{\mathbf{M}}_k)$ (per channel). This is saying that with each use of the big channel, Alice is able to get $\min_{k\in[K]} \mathrm{rk}_q(\overline{\mathbf{M}}_k)z$ sums in $\mathbb{F}_d$. Recall that each use of the big channel corresponds to the use of quantum subsystem $\mathcal{Q}_{t,s}$ with its dimension equal to $\delta_{t,s} = q^{N_{t,s}}$ for all $t \in [T], s \in \mathcal{E}(t)$. Since $q = d^z$, it follows that $\log_d \delta_{t,s} = N_{t,s}z$ and that the following set of tuples are feasible,

$$\text{closure}\left\{ \boldsymbol{\Delta} \in \mathbb{R}_+^{\Gamma} \;\middle|\; \begin{array}{l} N_{t,s} \in \mathbb{N}, \forall t \in [T], s \in \mathcal{E}(t), \\ \Delta_{t,s} \geq N_{t,s} / \min_{k\in[K]} \mathrm{rk}(\overline{\mathbf{M}}_k), t \in [T], s \in \mathcal{E}(t) \end{array} \right\}. \tag{54}$$

For fixed $N_{t,s}, t \in [T], s \in \mathcal{E}(t)$, we would like $\min_{k\in[K]} \mathrm{rk}(\overline{\mathbf{M}}_k)$ to attain its largest possible value to obtain the largest feasible set. According to (53), it suffices to let $\mathbf{M}_{t,k}$ have full rank for all $t \in [T], k \in [K]$. This can be done by letting $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_T$ be 'Half-MDS', which is defined as follows.

Definition 1 (Half-MDS). *Say $\mathbf{M} = [\mathbf{M}^l, \mathbf{M}^r] \in \mathbb{F}_q^{N\times 2N}$ is the transfer matrix of an N-sum box operating in $\mathbb{F}_q$. $\mathbf{M}^l \in \mathbb{F}_q^{N\times N}$ denotes the left half and $\mathbf{M}^r \in \mathbb{F}_q^{N\times N}$ denotes the right half. Let $i_1, i_2, \cdots, i_n \in \mathbb{N}$ be $n \leq N$ distinct indices not greater than N. We say $\mathbf{M}$ is half-MDS if for all such indices,*

$$\mathrm{rk}\big(\big[\mathbf{M}^l_{(i_1,\cdots,i_n)}, \mathbf{M}^r_{(i_1,\cdots,i_n)}\big]\big) = \min\{2n, N\},$$

where $\mathbf{M}_{(i_1,i_2,\cdots,i_n)}$ denotes the $N \times n$ submatrix of $\mathbf{M}$ comprised of the $(i_1, i_2, \cdots, i_n)^{th}$ columns of $\mathbf{M}$. As an example, consider feasible transfer matrices for 2-sum boxes,

$$\mathbf{M}_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \;\; \mathbf{M}_2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \tag{55}$$

Note that $\mathbf{M}_1$ is half-MDS while $\mathbf{M}_2$ is not. The submatrix comprised of the 2^{nd} and 4^{th} columns of $\mathbf{M}_2$ has rank $1 < 2$.

Lemma 4 (Half-MDS N-sum box). *If $q \geq N$, then there exists an N-sum box with transfer matrix $\mathbf{M} \in \mathbb{F}_q^{N\times 2N}$ that is half-MDS.*

The proof of Lemma 4 is presented in Appendix H.

Recall that we are free to choose z. Therefore, if we choose $z > \log_d(\max_{t\in[T]} N_t) \implies q = d^z > \max_{t\in[T]} N_t$, then $\overline{\mathbf{M}}_t$ can be made half-MDS for all $t \in [T]$. Now that $\mathbf{M}_t$ is half-MDS, the rank of $\mathbf{M}_{t,k}$ is equal to $\min(N_t, \sum_{s\in\mathcal{E}(t)\cap\mathcal{W}(k)} 2N_{t,s})$. Due to the diagonal block structure of $\overline{\mathbf{M}}_k$, the rank of $\overline{\mathbf{M}}_k$ is equal to $\sum_{t\in[T]} \min\left(N_t, \sum_{s\in\mathcal{E}(t)\cap\mathcal{W}(k)} 2N_{t,s}\right)$. Plugging this value into (54) we further obtain that the following set of tuples are feasible,

$$\text{closure}\left\{ \boldsymbol{\Delta} \in \mathbb{R}_+^{\Gamma} \;\middle|\; \begin{array}{l} N_{t,s} \in \mathbb{N}, \forall t \in [T], s \in \mathcal{E}(t), \\ R \in \mathbb{N}, \\ \sum_{t\in[T]} \min\left(N_t, \sum_{s\in\mathcal{E}(t)\cap\mathcal{W}(k)} 2N_{t,s}\right) \geq R, \; \forall k \in [K], \\ \Delta_{t,s} \geq N_{t,s}/R, t \in [T], s \in \mathcal{E}(t). \end{array} \right\} \tag{56}$$

$$= \left\{ \boldsymbol{\Delta} \in \mathbb{R}_+^{\Gamma} \;\middle|\; \sum_{t\in[T]} \min\left(\sum_{s\in\mathcal{E}(t)} \Delta_{t,s}, \sum_{s\in\mathcal{E}(t)\cap\mathcal{W}(k)} 2\Delta_{t,s}\right) \geq 1, \; \forall k \in [K] \right\}, \tag{57}$$

which is the same as the region $\mathcal{D}$ specified in Theorem 1.

Remark 1. *We remark that the initial quantum state that is used when constructing a half-MDS N-sum box is an absolutely maximally entangled (AME) state [16]. Specifically, say a quantum system $\mathcal{Q}$ of N qudits is in a pure state $|\psi\rangle$. Then the state is absolutely maximally entangled if for any partition of the system into $\mathcal{Q}_A$ with $k \leq \lfloor N/2 \rfloor$-qudits and $\mathcal{Q}_B$ with the remaining $N-k$ qudits, we have $S(\rho^A) = S(\rho^B) = k$ (dits), where ρ^A and ρ^B denote the reduced density matrices of $\mathcal{Q}_A$ and $\mathcal{Q}_B$, respectively. $S(\cdot)$ denotes the von Neumann entropy. According to [36], the generator matrix of the stabilizer code when constructing the box with the transfer matrix $\mathbf{M} = [\mathbf{M}_l, \mathbf{M}_r]$ is $\mathbf{G} = \begin{pmatrix} -\mathbf{M}_r^T \\ \mathbf{M}_l^T \end{pmatrix}$. Denote the initial state for this maximal stabilizer code as $|\mathbf{G}\rangle$. [62] shows that $|\mathbf{G}\rangle$ is equivalent to a graph state under local Clifford operations. [17] then shows that the graph state is maximally entangled if $\mathbf{G}^T$ is the generator matrix of an $[[n, k, d]]$ quantum MDS (stabilizer) code with $n = 2k$. This is satisfied because $\mathbf{M}$ is half-MDS, and so is $\mathbf{G}^T$. Since local Clifford operations do not affect the property of AME, we conclude that the initial state corresponding to the box $\mathbf{M}$ is an AME state.*

9 Conclusion

A sharp capacity characterization for the Σ-QMAC bodes well for future generalizations, that include in particular, the Linear Computation QMAC (LC-QMAC). As the quantum extension of the LC-MAC, which is the counterpart of the LCBC (linear computation broadcast) problem studied in [63, 64], the LC-QMAC assumes S servers, K data streams of $\mathbb{F}_d$ symbols, and a user (Alice) who wants to compute an *arbitrary $\mathbb{F}_d$ linear function* of the data streams. For example, with data streams $\mathsf{W}_1, \mathsf{W}_2, \cdots, \mathsf{W}_K$ represented as vectors over $\mathbb{F}_d$, Alice wants to compute $\mathsf{F} = V_1\mathsf{W}_1 + V_2\mathsf{W}_2 + \cdots + V_K\mathsf{W}_K$ for arbitrary linear transformations (matrices) $V_1, V_2, \cdots, V_K$ that are specified by the problem. Note that if $V_1, V_2, \cdots, V_K$ are invertible square matrices then the problem reduces to the Σ-QMAC, whose capacity is found in this work. This is because without loss of generality each $V_i\mathsf{W}_i$ can be defined to be a data stream $\widetilde{\mathsf{W}}_i$ over an extension field, leaving Alice only with the task of computing the sum, i.e., the Σ-QMAC setting. The general LC-QMAC setting, however, allows arbitrary matrices $V_1, V_2, \cdots, V_K$. In addition, the LC-QMAC specification includes arbitrary side-information at Alice of the form $\mathsf{F}' = V_1'\mathsf{W}_1 + V_2'\mathsf{W}_2 + \cdots + V_K'\mathsf{W}_K$, which can be quite useful for improving the communication efficiency of linear computation. Furthermore, the LC-QMAC allows the data available to each server to be arbitrary linear functions of the data streams, i.e., Server s has data of the form $U_{1s}'\mathsf{W}_1 + U_{2s}'\mathsf{W}_2 + \cdots + U_{Ks}'\mathsf{W}_K$, which adds another layer of both conceptual and combinatorial complexity. Indeed the capacity of even the classical LCMAC setting is not yet fully known. For example, consider an LC-QMAC setting over $\mathbb{F}_3$, with $K = 2$ data streams A, B, and 3 servers that have $(\mathsf{A}), (\mathsf{B}), (\mathsf{A} + 2\mathsf{B})$ respectively. Say Alice has no side-information and only wants to compute $\mathsf{A} + \mathsf{B}$. The capacity of this LC-MAC is not known to the best of our knowledge, but the corresponding quantum setting is trivial, i.e., the rate $R = 1$ qudit/dit is achieved simply if any two servers apply the 2-sum protocol. Similarly, there are QPIR settings where the capacity is known, while the corresponding classical cases remains open [37–40]. Thus, quantum settings can be tractable even when their classical counterparts are not. The sufficiency of the N-sum box abstraction for the LC-QMAC is an especially intriguing question. Aside from the LC-QMAC, generalizations in other directions, e.g., towards noisy quantum channels and correlated inputs as in [50–52], other forms of decoding locality restrictions as in [65], and to non-linear computations as in [24, 66] are also of interest.

Acknowledgment

The authors gratefully acknowledge helpful discussions with Matteo Allaix from Aalto University and Yuxiang Lu from University of California Irvine.

A Arbitrarily large DSC Gain over QMAC

Here we show that the DSC gain in the QMAC for certain partial function computations can be arbitrarily large by providing an example. The construction of this example largely relies on a bound of the chromatic number of the power of a family of graphs (referred to as the quarter-orthogonality graphs) presented in [67] that is based on earlier results in zero-error information theory such as [68,69], together with insights from quantum communication complexity literature such as [22,23].

Let $\kappa = 4p^r$ for an odd prime p and a positive integer r. For two vectors v_1, v_2 with the same length, let $h(v_1, v_2)$ denote the Hamming distance between them, i.e., the number of positions where their elements are distinct. Let A, B be data streams with realizations in $\{+1, -1\}^\kappa$ such that $h(\mathsf{A}, \mathsf{B}) \in \{0, \kappa/2\}$. There are 2 servers in the QMAC. Server A knows only A and Server B knows only B. The function to compute at Alice is defined as $\mathsf{F} = h(\mathsf{A}, \mathsf{B})$. Note that due to the dependence between A and B, the function to compute has binary output.

Let us first consider the capacity of the fully-unentangled case, C^{unent}. Denote by $\mathcal{Q}_A$ a quantum system of dimension δ_A that is sent from Server A and by $\mathcal{Q}_B$ a quantum system of dimension δ_B that is sent from Server B. Since our QMAC formulation does not allow the POVMs to depend on the data, and since there is no entanglement established between the two servers, without loss of generality, consider that Server A sends to Alice a state that is picked from a set of orthogonal states $\{|a\rangle_A\}_{a\in[\delta_A]}$ and that Server B sends to Alice a state from a set of orthogonal states $\{|b\rangle_B\}_{b\in[\delta_B]}$, which let Alice to perfectly recover (a, b). Suppose a genie provided Alice with $\mathsf{A}^{[L]}$. She would still need to recover $\mathsf{F}^{[L]}$ from $(b, \mathsf{A}^{[L]})$. Let $G_\kappa(V, E)$ be the *orthogonality graph* with vertices V uniquely mapping to $\{+1, -1\}^\kappa$ and for $v_1 \neq v_2 \in V$, $(v_1, v_2) \in E$ if and only if $h(v_1, v_2) = \kappa/2$. Let $H_{\kappa-1}(V', E')$ be the *quarter-orthogonality graph* with vertices V' uniquely mapping to the vectors in $\{+1, -1\}^{\kappa-1}$ that have an even number of "-1" entries, and for $v'_1 \neq v'_2 \in V'$, $(v'_1, v'_2) \in E'$ if and only if $h(v'_1, v'_2) = \kappa/2$. The quarter-orthogonality graph $H_{\kappa-1}$ is a subgraph of the orthogonality graph G_κ [67]. Similar to the reasoning in [68], $\delta_B \geq \chi(G_\kappa^{\boxtimes L})$ where $\boxtimes$ denotes the strong product of graphs[10] and $\chi(\cdot)$ denotes the chromatic number. [67, Cor. 5.8] shows that $\chi(H_{\kappa-1}^{\boxtimes L}) \geq 2^{(0.154\kappa - 1.154)L}$. Since $H_{\kappa-1}$ is a subgraph of G_κ, $H_{\kappa-1}^{\boxtimes L}$ is a subgraph of $G_\kappa^{\boxtimes L}$. It follows that $\chi(G_\kappa^{\boxtimes L}) \geq 2^{(0.154\kappa - 1.154)L}$ as the chromatic number of a graph cannot be less than the chromatic number of its subgraph. Thus, we obtain that $\log_2 \delta_B / L \geq 0.154\kappa - 1.154$. Due to symmetry between the two servers, we obtain that $C^{\text{unent}} \leq \sup_{L\to\infty} \frac{L}{\log_2 \delta_A + \log_2 \delta_B} \leq \frac{1/2}{0.154\kappa - 1.154}$ (computations per qubit).

Next, we show that the fully-entangled capacity $C^{\text{fullent}} \geq \frac{1/2}{\log_2 \kappa}$. Thus, the DSC gain is at least $\frac{0.154\kappa - 1.154}{\log_2 \kappa}$, which can be made arbitrary large by choosing κ large enough. The scheme

[10] For graphs G and H with respective vertex sets $V(G)$ and $V(H)$, define $G \boxtimes H$ as the strong product of G and H such that the vertex set of $G \boxtimes H$ is the Cartesian product $V(G) \times V(H)$; and distinct vertices (u, u') and (v, v') are adjacent in $G \boxtimes H$ if and only if: $u = v$ and u' is adjacent to v', or $u' = v'$ and u is adjacent to v, or u is adjacent to v and u' is adjacent to v'. $G^{\boxtimes L}$ is then defined as $\underbrace{G \boxtimes G \boxtimes \cdots \boxtimes G}_{L}$.

that achieves $\frac{1/2}{\log_2 \kappa}$ only needs batch size $L = 1$. The scheme is a generalization of the scheme in the problem referred to as the distributed Deutsch-Jozsa problem [22, 23] where $p = 2$ (but here we need p to be an odd prime). Let Server A and Server B share an entangled state $|\text{GHZ}\rangle = \frac{1}{\sqrt{\kappa}} \sum_{x\in[\kappa]} |xx\rangle$. Let $U_{\mathsf{A}} = \text{diag}(\mathsf{A})$ be a $\kappa \times \kappa$ diagonal matrix with the elements of vector A on the main diagonal. Similarly let $U_{\mathsf{B}} = \text{diag}(\mathsf{B})$. Note that U_{A} and U_{B} are unitary matrices. Let Server A apply the unitary operator U_{A} and Server B apply the unitary operator U_{B} to their respective quantum subsystems. Then the resulting state is $(U_{\mathsf{A}} \otimes U_{\mathsf{B}}) |\text{GHZ}\rangle = \frac{1}{\sqrt{\kappa}} \text{vec}(U_{\mathsf{B}}^T U_{\mathsf{A}})$, where $\text{vec}(\cdot)$ denotes the column-major vectorization function[11]. Note that we used the identity $\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B)$ [70] along with the fact that $|\text{GHZ}\rangle = \frac{1}{\sqrt{\kappa}} \text{vec}(\mathbf{I}_\kappa)$. Alice measures the quantum system by a PVM with two projectors $P_1 = |\text{GHZ}\rangle \langle\text{GHZ}|$ and $P_2 = \mathbf{I}_\kappa - P_1$. The measurement result being 1 (associated with P_1) has probability $\text{tr}(U_{\mathsf{B}}^T U_{\mathsf{A}})/\kappa = \mathsf{B}^T\mathsf{A}/\kappa$, which is equal to 1 if $h(\mathsf{A}, \mathsf{B}) = 0$, and equal to 0 if $h(\mathsf{A}, \mathsf{B}) = \kappa/2$. This means that Alice is able to distinguish the two possibilities of $h(\mathsf{A}, \mathsf{B})$ with certainty.

B Proof of Corollary 3

Let $\mathcal{D}_{1/2}^{\text{unent}} \triangleq \{(\Delta_1/2, \cdots, \Delta_S/2) \mid (\Delta_1, \cdots, \Delta_S) \in \mathcal{D}^{\text{unent}}\}$ and $\mathcal{D}_1 \triangleq \left\{(\Delta_1, \cdots, \Delta_S) \mid \sum_{s\in[S]} \Delta_s \geq 1\right\}$. Then Corollaries 1 and 2 together imply that

$$\mathcal{D}^{\text{fullent}} = \mathcal{D}_{1/2}^{\text{unent}} \cap \mathcal{D}_1. \tag{58}$$

Let $(\Delta_1^*, \cdots, \Delta_S^*)$ be a solution of $\arg\min_{(\Delta_1,\cdots,\Delta_S)\in\mathcal{D}^{\text{unent}}} \sum_{s\in[S]} \Delta_s$. It follows that $(\Delta_1^*/2, \cdots, \Delta_S^*/2)$ is a solution of $\arg\min_{(\Delta_1,\cdots,\Delta_S)\in\mathcal{D}_{1/2}^{\text{unent}}} \sum_{s\in[S]} \Delta_s$. Consider two cases.

1. If $\sum_{s\in[S]} \Delta_s^*/2 \geq 1$, i.e., $C^{\text{unent}} \leq 1/2$, then $(\Delta_1^*/2, \cdots, \Delta_S^*/2) \in \mathcal{D}_1$ and thus $(\Delta_1^*/2, \cdots, \Delta_S^*/2) \in \mathcal{D}^{\text{fullent}}$. It follows that $(\Delta_1^*/2, \cdots, \Delta_S^*/2)$ is a solution of $\arg\min_{(\Delta_1,\cdots,\Delta_S)\in\mathcal{D}^{\text{fullent}}} \sum_{s\in[S]} \Delta_s$. This implies that $C^{\text{fullent}} = (\sum_{s\in[S]} \Delta_s^*/2)^{-1} = 2(\sum_{s\in[S]} \Delta_s^*)^{-1} = 2C^{\text{unent}}$.
2. Otherwise, if $\sum_{s\in[S]} \Delta_s^*/2 < 1$, i.e., $C^{\text{unent}} > 1/2$, there exists $(\Delta_1', \cdots, \Delta_S')$ such that $\Delta_s' \geq \Delta_s^*/2, \forall s \in [S]$ and $\sum_{s\in[S]} \Delta_s' = 1$. Note that $(\Delta_1', \cdots, \Delta_S') \in \mathcal{D}_{1/2}^{\text{unent}}$ by the definition of $\mathcal{D}^{\text{unent}}$ (Corollary 2) and the definition of $\mathcal{D}_{1/2}^{\text{unent}}$. Since $\sum_{s\in[S]} \Delta_s' = 1$, we have $(\Delta_1', \cdots, \Delta_S') \in \mathcal{D}_1$ and therefore $(\Delta_1', \cdots, \Delta_S') \in \mathcal{D}^{\text{fullent}}$. This implies that $C^{\text{fullent}} \geq 1$ (and thus $C^{\text{fullent}} = 1$ as C^{fullent} is also upper bounded by 1).

Combining the two cases, we have $C^{\text{fullent}} = \min(1, 2C^{\text{unent}})$ and thus $C^{\text{fullent}}/C^{\text{unent}} = \min(2, 1/C^{\text{unent}})$.

C Proof of Corollary 4

C.1 Proof of (16)

Recall that for the symmetric settings, $K = \binom{S}{\alpha}, T = \binom{S}{\beta}$ and $\mathcal{W} : [K] \leftrightarrow \binom{[S]}{\alpha}$, $\mathcal{E} : [T] \leftrightarrow \binom{[S]}{\beta}$. We want to find

$$F^* \triangleq \min_{\mathbf{\Delta}\in\mathcal{D}} \sum_{t\in[T]} \sum_{s\in\mathcal{E}(t)} \Delta_{t,s}, \tag{59}$$

[11] $\text{vec}(A) \triangleq [a_{1,1}, \cdots, a_{m,1}, a_{1,2}, \cdots, a_{m,2}, \cdots, a_{1,n}, \cdots, a_{m,n}]^T$, where $a_{i,j}$ represents the element in the i^{th} row and j^{th} column of A.

where the feasible region $\mathcal{D}$ here is determined by Theorem 1 for the symmetric data replication and entanglement distribution maps. $C_\alpha^{(\beta)}$ immediately follows as $1/F^*$. Due to symmetry, the minimal value of $\sum_{t\in[T]}\sum_{s\in\mathcal{E}(t)}\Delta_{t,s}$ is achieved by $\Delta_{t,s}=\Delta_o\in\mathbb{R}_+, \forall t\in[T], s\in\mathcal{E}(t)$. Again by symmetry, the K conditions in $\mathcal{D}$ are identical in the form $f(\alpha,\beta)\Delta_o\geq 1$, where $f(\alpha,\beta)\in\mathbb{Z}^+$ is a function of (α,β). Next we derive $f(\alpha,\beta)$. Consider the value k such that $\mathcal{W}(k)=\{1,2,\cdots,\alpha\}=[\alpha]$. We have

$$f(\alpha,\beta)=\sum_{t\in[T]}\min\left(|\mathcal{E}(t)|,\ 2|\mathcal{E}(t)\cap[\alpha]|\right) \tag{60}$$

$$=\sum_{\mathcal{B}\subset\binom{[S]}{\beta}}\min\left(\beta,\ 2|\mathcal{B}\cap[\alpha]|\right) \tag{61}$$

$$=\sum_{\gamma=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)}\min(\beta,\ 2\gamma)\cdot\left|\left\{\mathcal{B}\in\tbinom{[S]}{\beta}\mid\mathcal{B}\cap[\alpha]=\gamma\right\}\right| \tag{62}$$

$$=\sum_{\gamma=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)}\min(\beta,\ 2\gamma)\cdot\binom{\alpha}{\gamma}\cdot\binom{S-\alpha}{\beta-\gamma}. \tag{63}$$

It follows that $F^*=\frac{\beta T}{f(\alpha,\beta)}$ and therefore $C_\alpha^{(\beta)}=1/F^*=\frac{f(\alpha,\beta)}{\beta T}$.

C.2 Proof of (17)

We can rewrite $C_\alpha^{(\beta)}$ as

$$C_\alpha^{(\beta)}=\frac{1}{\beta T}\sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)}\left(2\gamma-(2\gamma-\beta)^+\right)\cdot\binom{\alpha}{\gamma}\cdot\binom{S-\alpha}{\beta-\gamma} \tag{64}$$

$$=\frac{2}{\beta T}\sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)}\gamma\cdot\binom{\alpha}{\gamma}\cdot\binom{S-\alpha}{\beta-\gamma}-\frac{1}{\beta T}\sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)}(2\gamma-\beta)^+\cdot\binom{\alpha}{\gamma}\cdot\binom{S-\alpha}{\beta-\gamma} \tag{65}$$

$$=\frac{2\alpha}{\beta T}\sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)}\binom{\alpha-1}{\gamma-1}\cdot\binom{S-\alpha}{\beta-\gamma}-\frac{1}{\beta T}\sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)}(2\gamma-\beta)^+\cdot\binom{\alpha}{\gamma}\cdot\binom{S-\alpha}{\beta-\gamma} \tag{66}$$

$$=\frac{2\alpha}{\beta T}\cdot\binom{S-1}{\beta-1}-\frac{1}{\beta T}\sum_{r=\max(\alpha+\beta-S,\lceil\beta/2\rceil)}^{\min(\alpha,\beta)}(2\gamma-\beta)\cdot\binom{\alpha}{\gamma}\cdot\binom{S-\alpha}{\beta-\gamma} \tag{67}$$

$$=\frac{2\alpha}{S}-\frac{1}{\beta T}\sum_{r=\max(\alpha+\beta-S,\lceil\beta/2\rceil)}^{\min(\alpha,\beta)}(2\gamma-\beta)\cdot\binom{\alpha}{\gamma}\cdot\binom{S-\alpha}{\beta-\gamma} \tag{68}$$

C.3 Proof of (18)

Alternatively, we can rewrite $C_\alpha^{(\beta)}$ as

$$C_\alpha^{(\beta)}=\frac{1}{\beta T}\sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)}\left(\beta-(\beta-2\gamma)^+\right)\cdot\binom{\alpha}{\gamma}\cdot\binom{S-\alpha}{\beta-\gamma} \tag{69}$$

$$= \frac{1}{T} \sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)} \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma} - \frac{1}{\beta T} \sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\beta)} (\beta-2\gamma)^+ \cdot \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma} \tag{70}$$

$$= \frac{1}{T}\binom{S}{\beta} - \frac{1}{\beta T} \sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\lfloor\beta/2\rfloor)} (\beta-2\gamma) \cdot \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma} \tag{71}$$

$$= 1 - \frac{1}{\beta T} \sum_{r=(\alpha+\beta-S)^+}^{\min(\alpha,\lfloor\beta/2\rfloor)} (\beta-2\gamma) \cdot \binom{\alpha}{\gamma} \cdot \binom{S-\alpha}{\beta-\gamma} \tag{72}$$

C.4 Proof of (29)

The case when $\alpha = S$ is trivial. For $\alpha \leq \lfloor S/2 \rfloor$, let us make use of (17). It is not difficult to obtain that $C_\alpha^{(\beta)} = \frac{2\alpha}{S}$ when $\beta \geq 2\alpha$, as $0 \leq 2\gamma - \beta \leq 2\alpha - \beta \leq 0$ for $\lceil \beta/2 \rceil \leq \gamma \leq \alpha$, so either $2\gamma - \beta = 0$ or γ does not take any value in the summation term. On the other hand, if $\beta < 2\alpha$, we have $\alpha \geq \lceil \beta/2 \rceil$ and thus $\min(\alpha, \beta) \geq \max(\alpha + \beta - S, \lceil \beta/2 \rceil)$, so γ must take at least one value in the summation. It follows that $C_\alpha^{(\beta)} < \frac{2\alpha}{S}$ because $2\min(\alpha, \beta) - \beta > 0$. Since $C_\alpha^{(S)} = \frac{2\alpha}{S}$ for $\alpha \leq \lfloor S/2 \rfloor$, we obtain that $\beta^* = 2\alpha$ for these cases.

For $\lceil S/2 \rceil \leq \alpha \leq S-1$, let us make use of (18). It is not difficult to obtain that $C_\alpha^{(\beta)} = 1$ when $\beta \geq 2(S-\alpha)$, as $0 \leq \beta - 2\gamma \leq 2(S-\alpha) - \beta \leq 0$ for $\lfloor \beta/2 \rfloor \leq \gamma \leq 2(S-\alpha)$, so either $\beta - 2\gamma = 0$ or γ does not take any value in the summation term. On the other hand, if $\beta < 2(S-\alpha)$, we have $\lceil \beta/2 \rceil \leq S - \alpha \implies \lfloor \beta/2 \rfloor \geq \alpha + \beta - S$ and thus $\min(\alpha, \lfloor \beta/2 \rfloor) \geq (\alpha + \beta - S)^+$, so γ must take at least one value in the summation. It follows that $C_\alpha^{(\beta)} < 1$ because $\beta - 2(\alpha + \beta - S)^+ > 0$. Since $C_\alpha^{(S)} = 1$ for $\lceil S/2 \rceil \leq \alpha \leq S-1$, we obtain that $\beta^* = 2(S-\alpha)$ for these cases.

D Proof of Corollary 5

First let us note that there are two Σ-QMAC problems involved in the corollary, summarized as follows.

1. The original problem $\mathcal{P}$ has data replication map $\mathcal{W}$, S servers and K data streams. We refer to the S servers in $\mathcal{P}$ by $\mathcal{S}_i$ for $i \in [S]$. We are interested in the 2-party-entangled capacity $C^{(2)}(\mathcal{W})$.

2. The hypothetical problem $\widetilde{\mathcal{P}}$ has data replication map $\widetilde{\mathcal{W}}$, $\binom{S}{2}$ servers and the same K data streams as in $\mathcal{P}$. We refer to the $\binom{S}{2}$ servers in $\widetilde{\mathcal{P}}$ by $\mathcal{S}_{\{i,j\}}$ for $\{i,j\} \in \binom{[S]}{2}$. $\mathcal{S}_{\{i,j\}}$ has the access to the data streams that are available to either $\mathcal{S}_i$ or $\mathcal{S}_j$ in $\mathcal{P}$. We are interested in the fully-unentangled capacity $C^{\text{unent}}(\widetilde{\mathcal{W}})$.

Our goal is to prove that $C^{(2)}(\mathcal{W}) = C^{\text{unent}}(\widetilde{\mathcal{W}})$. The proposition is comprised of two bounds, $C^{(2)}(\mathcal{W}) \leq \widetilde{C}^o(\widetilde{\mathcal{W}})$ and $C^{(2)}(\mathcal{W}) \geq C^{\text{unent}}(\widetilde{\mathcal{W}})$. To prove $C^{(2)}(\mathcal{W}) \leq C^{\text{unent}}(\widetilde{\mathcal{W}})$, consider any rate R that is achievable in the problem $\mathcal{P}$ with only bipartite entanglement, i.e., any clique contains at most 2 servers. The output state corresponds to any clique $\{i,j\} \in \binom{[S]}{2}$ of this scheme can always be generated by the server $\mathcal{S}_{\{i,j\}}$ in the problem $\widetilde{\mathcal{P}}$, since this server has the access to all data streams that are available to either $\mathcal{S}_i$ or $\mathcal{S}_j$ in the problem $\mathcal{P}$. In the problem $\widetilde{\mathcal{P}}$, a scheme can

let $\mathcal{S}_{\{i,j\}}$ directly transmit this state to Alice. Therefore, R must be achievable in $\widetilde{\mathcal{P}}$ as well, which shows that $C^{(2)}(\mathcal{W}) \leq C^{\text{unent}}(\widetilde{\mathcal{W}})$.

To prove $C^{(2)}(\mathcal{W}) \geq C^{\text{unent}}(\widetilde{\mathcal{W}})$, we need an intermediate result of Theorem 1 that coding based on the N-sum box abstraction is optimal in every case. For the unentangled case, this means that each server is simply treating qudits as classical dits. Additionlly, Lemma 3 implies that the optimal scheme is also linear. With this knowledge, given that R is achievable in the problem $\widetilde{\mathcal{P}}$, let us consider a linear scheme that achieves this R, where the coders $V_{\{i,j\}} \in \mathbb{F}_d^{N_{\{i,j\}} \times L}$ at the server $\mathcal{S}_{\{i,j\}}$ takes as input L symbols of all its accessible data streams, i.e., $\mathsf{W}_k, k \in \widetilde{\mathcal{W}}(\{i,j\})$, and outputs a vector $Y_{\{i,j\}} \in \mathbb{F}_d^{N_{\{i,j\}} \times 1}$. In this linear scheme, symbols are considered in the field $\mathbb{F}_d$. Say $\mathcal{S}_{\{i,j\}}$ transmits $N_{\{i,j\}}$ qudits to inform Alice of $Y_{\{i,j\}}$. Upon receiving the $\sum_{\{i,j\} \in \binom{[S]}{2}} N_{\{i,j\}} \triangleq N$ qudits, Alice computes L instances of sum. The scheme satisfies that $\frac{L}{N} \geq R$ by definition. What we will do next is to convert this scheme to a scheme in the problem $\mathcal{P}$ that achieves the same rate with only the use of the 2-sum protocol. Recall that the 2-sum protocol behaves the same as an $\mathbb{F}_d$ additive channel, in the way that the receiver is able to get one sum of the two $\mathbb{F}_d$ inputs from the two transmitters with the cost of one qudit on average. We also point out that each use of the 2-sum protocol is equivalent to use such an additive channel twice, i.e., the receiver gets two dimensions of the sums at a cost of 2 qudits.

Now let us look at the original problem $\mathcal{P}$. Let us construct a scheme with batch size $2L$, so that Alice is able to compute $2L$ instances of the sum by using the 2-sum protocol N times, at a cost of $2N$ qudits, thus achieving the same rate as in the problem $\widetilde{\mathcal{P}}$. Recall that $Y_{\{i,j\}}$ denotes the transmission from the server $\mathcal{S}_{\{i,j\}}$ in the problem $\widetilde{\mathcal{P}}$. Since $Y_{\{i,j\}} \in \mathbb{F}_d^{N_{\{i,j\}} \times 1}$ is a linear function of the data streams that are known to either $\mathcal{S}_i$ or $\mathcal{S}_j$ in the problem $\mathcal{P}$, we can represent $Y_{\{i,j\}} = Y_{\{i,j\},i} + Y_{\{i,j\},j}$, where $Y_{\{i,j\},i} \in \mathbb{F}_d^{N_{\{i,j\}} \times 1}$ can be computed by $\mathcal{S}_i$, and $Y_{\{i,j\},j} \in \mathbb{F}_d^{N_{\{i,j\}} \times 1}$ can be computed by Server $\mathcal{S}_j$ in the problem $\mathcal{P}$. Therefore, with $N_{\{i,j\}}$ use of the binary additive channel in $\mathbb{F}_d$ between $\mathcal{S}_i$ and $\mathcal{S}_j$, the two servers can transmit the sum $Y_{\{i,j\}}$ to Alice. To apply the 2-sum protocol, we only need to consider two parallel instances of $Y_{\{i,j\}}$, so that with $N_{\{i,j\}}$ uses of the 2-sum protocol between $\mathcal{S}_i$ and $\mathcal{S}_j$, Alice is able to get 2 instances of $Y_{\{i,j\}}$. Taking all pairs of servers into account, for the converted scheme in the problem $\mathcal{P}$, the 2-sum protocol (in $\mathbb{F}_d$) is used $\sum_{\{i,j\} \in \binom{[S]}{2}} N_{\{i,j\}} = N$ times in total, allowing Alice to compute $2N$ instances of the sum in $\mathbb{F}_d$.

E Proof of Corollary 7

We show that for any data replication map $\mathcal{W}$, and entanglement distribution map $\mathcal{E}$ where $|\mathcal{E}(t)| = 3$ for some t, the clique $\mathcal{E}(t)$ can be replaced by three bipartite cliques (that contain only two servers) without decreasing the capacity. Formally, let $\mathcal{E}$ be such an entanglement distribution map and without loss of generality, $\mathcal{E}(1) = \{1,2,3\}$. Let $\widetilde{\mathcal{E}}$ be another entanglement distribution map such that $\widetilde{\mathcal{E}}(1) = \{2,3\}$, $\widetilde{\mathcal{E}}(2) = \{1,3\}$, $\widetilde{\mathcal{E}}(3) = \{1,2\}$, and $\widetilde{\mathcal{E}}(t) = \mathcal{E}(t-2)$ for $t \geq 4$. Let $\mathcal{P}, \widetilde{\mathcal{P}}$ denote the respective Σ-QMAC problems with entanglement distribution maps $\mathcal{E}$, $\widetilde{\mathcal{E}}$, and both problems have data replication map $\mathcal{W}$. For the problem $\mathcal{P}$, we use $\Delta_{t,s}$ to denote the (normalized) download cost associated with Clique $\mathcal{E}(t)$ and Server $s \in \mathcal{E}(t)$, and we use $\boldsymbol{\Delta}$ to denote the download cost tuple. To avoid confusion, in the problem $\widetilde{\mathcal{P}}$, we use $\widetilde{\Delta}_{t,s}$ to denote the download cost associated with Clique $\widetilde{\mathcal{E}}(t)$ and Server $s \in \widetilde{\mathcal{E}}(t)$, and we use $\widetilde{\boldsymbol{\Delta}}$ to denote the download cost

tuple. Since $\mathcal{E}(1) = \{1,2,3\}$ is our main focus, in the following we let $\Delta_s \triangleq \Delta_{1,s}, s \in \{1,2,3\}$ for brevity. Our goal is to show that $C(\mathcal{W}, \widetilde{\mathcal{E}}) \geq C(\mathcal{W}, \mathcal{E})$ and thus $C(\mathcal{W}, \widetilde{\mathcal{E}}) = C(\mathcal{W}, \mathcal{E})$ because any coding scheme allowed in the problem $\widetilde{\mathcal{P}}$ is also allowed in the problem $\mathcal{P}$.

Consider any feasible download tuple $\boldsymbol{\Delta}$ (in the feasible region implied by Theorem 1) for the problem $\mathcal{P}$. To focus on the clique $\mathcal{E}(1)$, Theorem 1 implies that $\boldsymbol{\Delta}$ is feasible if and only if,

$$\begin{cases} \min\{\Delta_1+\Delta_2+\Delta_3, 2\Delta_1\} \geq c_1 \\ \min\{\Delta_1+\Delta_2+\Delta_3, 2\Delta_2\} \geq c_2 \\ \min\{\Delta_1+\Delta_2+\Delta_3, 2\Delta_3\} \geq c_3 \\ \min\{\Delta_1+\Delta_2+\Delta_3, 2\Delta_1+2\Delta_2\} \geq c_{12} \\ \min\{\Delta_1+\Delta_2+\Delta_3, 2\Delta_1+2\Delta_3\} \geq c_{13} \\ \min\{\Delta_1+\Delta_2+\Delta_3, 2\Delta_2+2\Delta_3\} \geq c_{23} \\ \Delta_1+\Delta_2+\Delta_3 \geq c_{123} \end{cases} \tag{73}$$

where $c_1, c_2, \cdots, c_{123}$ are determined by $\mathcal{W}$ and $\left(\Delta_{t,s}\right)_{t\geq 2, s\in\mathcal{E}(t)}$, i.e., the (normalized) download costs associated with the other remaining cliques in $\mathcal{E}$.

Let us note that it is without loss of generality to consider such feasible tuples with $\Delta_i \leq \Delta_j + \Delta_k$ for $\{i,j,k\} \in \{\{1,2,3\},\{2,1,3\},\{3,1,2\}\}$ if we are only interested in their sum $\Delta_1+\Delta_2+\Delta_3$, because otherwise (say $\Delta_1 > \Delta_2+\Delta_3$) we can let

$$\begin{bmatrix} \Delta_1' \\ \Delta_2' \\ \Delta_3' \end{bmatrix} = \begin{bmatrix} (\Delta_1+\Delta_2+\Delta_3)/2 \\ (\Delta_1+\Delta_2-\Delta_3)/2 \\ \Delta_3 \end{bmatrix} \tag{74}$$

so that (73) is also satisfied if we replace $(\Delta_1, \Delta_2, \Delta_3)$ with $(\Delta_1', \Delta_2', \Delta_3')$. Note that $\Delta_1'+\Delta_2'+\Delta_3' = \Delta_1+\Delta_2+\Delta_3$ but now $\Delta_1' = \Delta_2' + \Delta_3'$.

Now let us study the problem $\widetilde{\mathcal{P}}$. Note that by definition, $\widetilde{\mathcal{E}}(t) = \mathcal{E}(t-2)$ for $t \geq 4$. Then by Theorem 1, the download cost tuple $\widetilde{\boldsymbol{\Delta}}$ is feasible if and only if

$$\begin{cases} \widetilde{\Delta}_{t,s} = \Delta_{t-2,s}, \forall t \geq 4, s \in \widetilde{\mathcal{E}}(t) \\ \min\{\widetilde{\Delta}_{2,1}+\widetilde{\Delta}_{2,3}, 2\widetilde{\Delta}_{2,1}\} + \min\{\widetilde{\Delta}_{3,1}+\widetilde{\Delta}_{3,2}, 2\widetilde{\Delta}_{3,1}\} \geq c_1 \\ \min\{\widetilde{\Delta}_{1,2}+\widetilde{\Delta}_{1,3}, 2\widetilde{\Delta}_{1,2}\} + \min\{\widetilde{\Delta}_{3,1}+\widetilde{\Delta}_{3,2}, 2\widetilde{\Delta}_{3,2}\} \geq c_2 \\ \min\{\widetilde{\Delta}_{1,2}+\widetilde{\Delta}_{1,3}, 2\widetilde{\Delta}_{1,3}\} + \min\{\widetilde{\Delta}_{2,1}+\widetilde{\Delta}_{2,3}, 2\widetilde{\Delta}_{2,3}\} \geq c_3 \\ \min\{\widetilde{\Delta}_{1,2}+\widetilde{\Delta}_{1,3}, 2\widetilde{\Delta}_{1,2}\} + \min\{\widetilde{\Delta}_{2,1}+\widetilde{\Delta}_{2,3}, 2\widetilde{\Delta}_{2,1}\} + \widetilde{\Delta}_{3,1}+\widetilde{\Delta}_{3,2} \geq c_{12} \\ \min\{\widetilde{\Delta}_{1,2}+\widetilde{\Delta}_{1,3}, 2\widetilde{\Delta}_{1,3}\} + \widetilde{\Delta}_{2,1}+\widetilde{\Delta}_{2,3} + \min\{\widetilde{\Delta}_{3,1}+\widetilde{\Delta}_{3,2}, 2\widetilde{\Delta}_{3,1}\} \geq c_{13} \\ \widetilde{\Delta}_{1,2}+\widetilde{\Delta}_{1,3} + \min\{\widetilde{\Delta}_{2,1}+\widetilde{\Delta}_{2,3}, 2\widetilde{\Delta}_{2,3}\} + \min\{\widetilde{\Delta}_{3,1}+\widetilde{\Delta}_{3,2}, 2\widetilde{\Delta}_{3,2}\} \geq c_{23} \\ \widetilde{\Delta}_{1,2}+\widetilde{\Delta}_{1,3}+\widetilde{\Delta}_{2,1}+\widetilde{\Delta}_{2,3}+\widetilde{\Delta}_{3,1}+\widetilde{\Delta}_{3,2} \geq c_{123} \end{cases} \tag{75}$$

where $c_1, c_2, \cdots, c_{123}$ are the same as those in (73). Now, consider the download cost tuple $\widetilde{\boldsymbol{\Delta}}$ in the problem $\widetilde{\mathcal{P}}$, such that

$$\begin{aligned} &\widetilde{\Delta}_{t,s} = \Delta_{t-2,s}, \forall t \geq 4, s \in \widetilde{\mathcal{E}}(t) \\ \text{and } &\begin{bmatrix} \widetilde{\Delta}_{1,2} \\ \widetilde{\Delta}_{2,1} \\ \widetilde{\Delta}_{3,1} \end{bmatrix} = \begin{bmatrix} \widetilde{\Delta}_{1,3} \\ \widetilde{\Delta}_{2,3} \\ \widetilde{\Delta}_{3,2} \end{bmatrix} = \begin{bmatrix} (\Delta_2+\Delta_3-\Delta_1)/2 \\ (\Delta_1+\Delta_3-\Delta_2)/2 \\ (\Delta_1+\Delta_2-\Delta_3)/2 \end{bmatrix}. \end{aligned} \tag{76}$$

Then it can be verified that the download cost tuple $\widetilde{\mathbf{\Delta}}$ is feasible in $\widetilde{\mathcal{P}}$ if $\mathbf{\Delta}$ is feasible in $\mathcal{P}$, because (75) is satisfied if (73) is satisfied. Therefore, the feasibility of $\mathbf{\Delta}$ in the problem $\mathcal{P}$ implies the feasibility of $\widetilde{\mathbf{\Delta}}$ in the problem $\widetilde{\mathcal{P}}$. Since $\widetilde{\Delta}_{1,2}+\widetilde{\Delta}_{1,3}+\widetilde{\Delta}_{2,1}+\widetilde{\Delta}_{2,3}+\widetilde{\Delta}_{3,1}+\widetilde{\Delta}_{3,2} = \Delta_1+\Delta_2+\Delta_3$ and $\sum_{t\geq 4, s\in\widetilde{\mathcal{E}}(t)} \widetilde{\Delta}_{t,s} = \sum_{t\geq 2, s\in\mathcal{E}(t)} \Delta_{t,s}$ by (76), $\mathbf{\Delta}$ and $\tilde{\mathbf{\Delta}}$ have the same (normalized) sum download costs. Since it holds for any feasible download cost tuple $\mathbf{\Delta}$ in the problem $\mathcal{P}$, it follows that $C(\mathcal{W},\mathcal{E}') \geq C(\mathcal{W},\mathcal{E})$.

F Proof of Corollary 8

When S is even, this corollary can be easily shown with the symmetric data replication maps as defined in Corollary 4. Consider the Σ-QMAC with the symmetric data replication maps with $\alpha = S/2$. Then (29) says that $\beta^* = S \implies C^{(S)}_{S/2} > C^{(S-1)}_{S/2}$.

When S is odd, let us consider the data replication map $\mathcal{W} = \binom{[S-1]}{S-2} \cup \{S\}$. Without loss of generality, say $\mathcal{W}(S) = \{S\}$. We first apply Corollary 1 to show that $C^{\text{fullent}}(\mathcal{W}) \geq \frac{2S-4}{2S-3}$. Let Δ_s be the normalized download cost from Server s for $s \in [S]$. Writing down the feasible region by Corollary 1 explicitly for this setting, we have

$$\mathcal{D}^{\text{fullent}}(\mathcal{W}) = \left\{ (\Delta_1,\cdots,\Delta_S) \in \mathbb{R}_+^S \;\middle|\; \begin{array}{l} \Delta_1+\Delta_2+\cdots+\Delta_S \geq 1 \\ 2(\Delta_1+\Delta_2+\cdots+\Delta_{S-1}+\Delta_{S-2}) \geq 1 \\ 2(\Delta_1+\Delta_2+\cdots+\Delta_{S-3}+\Delta_{S-1}) \geq 1 \\ \quad\vdots \\ 2(\Delta_2+\Delta_3+\cdots+\Delta_{S-2}+\Delta_{S-1}) \geq 1 \\ 2\Delta_S \geq 1 \end{array} \right\}.$$

It can be verified that

$$(\Delta_1,\cdots,\Delta_{S-1},\Delta_S) = \left(\underbrace{\frac{1}{2(S-2)},\cdots,\frac{1}{2(S-2)}}_{S-1}, \frac{1}{2} \right) \in \mathcal{D}^{\text{fullent}}(\mathcal{W}). \tag{77}$$

We thus obtain that $\Delta_1+\Delta_2+\cdots+\Delta_S = \frac{S-1}{2(S-2)} + \frac{1}{2} = \frac{2S-3}{2S-4}$. It then follows that $C^{\text{fullent}}(\mathcal{W}) = \left(\min_{(\Delta_1,\cdots,\Delta_S)\in\mathcal{D}^{\text{fullent}}(\mathcal{W})} \sum_{s\in[S]} \Delta_s\right)^{-1} \geq \frac{2S-4}{2S-3}$.

Next we show that $C^{(S-1)}(\mathcal{W}) \leq \frac{2S-5}{2S-4}$ by Theorem 1. Note that for the entanglement distribution map $\mathcal{E} = \binom{[S]}{S-1}$, the region specified in Theorem 1 contains $\Gamma = S(S-1)$ variables. This is because there are $T = S$ cliques and the size of each clique is $S-1$. Also note that for this setting we have $\mathcal{E}(i) \neq \mathcal{E}(j)$ for $i \neq j$. Therefore, whenever it is needed to explicitly identify the servers in a specified clique, we use $\Delta_{\mathcal{E}(t),s}$ to replace $\Delta_{t,s}$, so that it becomes clear which servers are contained in the clique. Also, let $\Delta_{\mathcal{E}(t)} \triangleq \sum_{s\in\mathcal{E}(t)} \Delta_{\mathcal{E}(t),s}$. By Theorem 1, the feasibility of $\mathbf{\Delta}$ implies that

$$\Delta_{\{1,2,\cdots,S-1\}} + \Delta_{\mathcal{W}(k)\cup\{S\}} + \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}\setminus\{\mathcal{W}(k)\}} \sum_{s\in(\mathcal{A}\cup\{S\})\cap\mathcal{W}(k)} 2\Delta_{\mathcal{A}\cup\{S\},s} \geq 1, \;\; \forall k\in[S-1], \tag{78}$$

which can be simplified as

$$\Delta_{\{1,2,\cdots,S-1\}} + \Delta_{\mathcal{B}\cup\{S\}} + \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}\setminus\{\mathcal{B}\}} \sum_{s\in\mathcal{A}\cap\mathcal{B}} 2\Delta_{\mathcal{A}\cup\{S\},s} \geq 1, \ \ \forall \mathcal{B} \in \binom{[S-1]}{S-2}, \tag{79}$$

where we use $\mathcal{B}$ to substitute $\mathcal{W}(k)$ and note that $\mathcal{W}(k) \cap \{S\} = \emptyset$ for any $k \in [S-1]$. For the last data stream, since $\mathcal{W}(S) = \{S\}$, data stream W_S is not available to any server in the clique $\{1, 2, \cdots, S-1\}$ and thus the feasibility of $\mathbf{\Delta}$ implies that

$$\sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} 2\Delta_{\mathcal{A}\cup\{S\},S} \geq 1. \tag{80}$$

Therefore,

$$\begin{aligned}
&2S-4 \\
&= (S-1) + (S-3) && (81)\\
&\leq \sum_{\mathcal{B}\in\binom{[S-1]}{S-2}} \left(\Delta_{\{1,2,\cdots,S-1\}} + \Delta_{\mathcal{B}\cup\{S\}} + \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}\setminus\{\mathcal{B}\}} \sum_{s\in\mathcal{A}\cap\mathcal{B}} 2\Delta_{\mathcal{A}\cup\{S\},s} \right) + \underbrace{(S-3) \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} 2\Delta_{\mathcal{A}\cup\{S\},S}}_{\Xi_1} && (82)\\
&= \underbrace{(S-1)\Delta_{\{1,2,\cdots,S-1\}} + \sum_{\mathcal{B}\in\binom{[S-1]}{S-2}} \Delta_{\mathcal{B}\cup\{S\}}}_{\Xi_2} + \sum_{\mathcal{B}\in\binom{[S-1]}{S-2}} \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}\setminus\{\mathcal{B}\}} \sum_{s\in\mathcal{A}\cap\mathcal{B}} 2\Delta_{\mathcal{A}\cup\{S\},s} + \Xi_1 && (83)\\
&= \Xi_1 + \Xi_2 + \sum_{\mathcal{B}\in\binom{[S-1]}{S-2}} \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} \sum_{s\in\mathcal{A}\cap\mathcal{B}} 2\Delta_{\mathcal{A}\cap\{S\},s} - \sum_{\mathcal{B}\in\binom{[S-2]}{S-1}} \sum_{s\in\mathcal{B}} 2\Delta_{\mathcal{B}\cup\{S\},s} && (84)\\
&= \Xi_1 + \Xi_2 + \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} \sum_{\mathcal{B}\in\binom{[S-1]}{S-2}} \sum_{s\in\mathcal{A}\cap\mathcal{B}} 2\Delta_{\mathcal{A}\cap\{S\},s} - \sum_{\mathcal{B}\in\binom{[S-2]}{S-1}} \sum_{s\in\mathcal{B}} 2\Delta_{\mathcal{B}\cup\{S\},s} && (85)\\
&= \Xi_1 + \Xi_2 + (S-2) \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} \sum_{s\in\mathcal{A}} 2\Delta_{\mathcal{A}\cup\{S\},s} - \sum_{\mathcal{B}\in\binom{[S-2]}{S-1}} \sum_{s\in\mathcal{B}} 2\Delta_{\mathcal{B}\cup\{S\},s} && (86)\\
&= \Xi_1 + \Xi_2 + (S-3) \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} \sum_{s\in\mathcal{A}} 2\Delta_{\mathcal{A}\cup\{S\},s} && (87)\\
&= \Xi_2 + (S-3) \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} \sum_{s\in\mathcal{A}\cup\{S\}} 2\Delta_{\mathcal{A}\cup\{S\},s} && (88)\\
&= \Xi_2 + (S-3) \sum_{\mathcal{A}\in\binom{[S-1]}{S-1}} 2\Delta_{\mathcal{A}\cup\{S\}} && (89)\\
&= (S-1)\Delta_{\{1,2,\cdots,S-1\}} + \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} \underbrace{\big(1+2(S-3)\big)}_{2S-5} \Delta_{\mathcal{A}\cup\{S\}} && (90)\\
&\leq \left(2S-5\right) \left(\Delta_{\{1,2,\cdots,S-1\}} + \sum_{\mathcal{A}\in\binom{[S-1]}{S-2}} \Delta_{\mathcal{A}\cup\{S\}} \right) && (91)\\
&= (2S-5) \sum_{t\in[T]} \Delta_{\mathcal{E}(t)} && (92)
\end{aligned}$$

$$= (2S-5) \sum_{t\in[T]} \sum_{s\in\mathcal{E}(t)} \Delta_{t,s} \tag{93}$$

Step (82) is by (79) and (80). To see Step (86), the term $2\Delta_{\mathcal{A}\cup\{S\},s}$ is counted $S-2$ times for any specified $\mathcal{A} \in \binom{[S-1]}{S-2}$ and $s \in \mathcal{A}$, because there is exactly one $\mathcal{B} \in \binom{[S-1]}{S-2}$ such that $s \notin \mathcal{A}\cap\mathcal{B}$, which is $\mathcal{B} = [S-1]\backslash\{s\}$. Therefore, there are $(S-2)$ $\mathcal{B} \in \binom{[S-1]}{S-2}$ for which $s \in \mathcal{A}\cap\mathcal{B}$. Step (91) is because $(S-1) < 2S-5$ and $\Delta_{\{1,2,\cdots,S-1\}} \geq 0$.

Since $C^{(S-1)}(\mathcal{W}) = \left(\min_{\boldsymbol{\Delta}\in\mathcal{D}} \sum_{t\in[T],s\in\mathcal{E}(t)} \Delta_{t,s}\right)^{-1}$, we conclude that $C^{(S-1)}(\mathcal{W}) \leq \frac{2S-5}{2S-4}$.

G Proof of Lemma 3

The converse is obvious, as when there is only one receiver, the capacity cannot exceed $\mathrm{rk}(\mathbf{H}_1)$. Next let us consider the achievability. We want to design a scheme with batch size L and N channel uses such that $L/N = \min_{k\in[K]} \mathrm{rk}_q(\mathbf{H}_k)$. Note that since we use the channel N times, we can consider the input $\widetilde{X} \in \mathbb{F}_{q^N}^{n\times 1}$ and $\widetilde{Y} \in \mathbb{F}_{q^N}^{m_k\times 1}$. Let L, N be such integers that $q^N > K \min_{k\in[K]} \mathrm{rk}(\mathbf{H}_k)$ and $L = N \min_{k\in[K]} \mathrm{rk}(\mathbf{H}_k)$, i.e., $L/N = \min_{k\in[K]} \mathrm{rk}(\mathbf{H}_k)$. Take L symbols from the data stream W and regard it as a vector $\mathbf{W} \in \mathbb{F}_{q^N}^{L/N\times 1}$. For each $k \in [K]$, since $\mathrm{rk}(\mathbf{H}_k) \geq L/N$, there exist matrices $\overline{\mathbf{U}}_k \in \mathbb{F}_{q^N}^{L/N\times m_k}$, $\overline{\mathbf{V}}_k \in \mathbb{F}_{q^N}^{n\times L/N}$ such that $\overline{\mathbf{U}}_k\mathbf{H}_k^T\overline{\mathbf{V}}_k = \mathbf{I}_{L/N}$. Now consider a matrix $\mathbf{V} \in \mathbb{F}_{q^N}^{n\times L/N}$ whose elements are variables in $\mathbb{F}_{q^N}$ with values yet to be determined. Note that $P_k \triangleq \det(\overline{\mathbf{U}}_k\mathbf{M}_k\mathbf{V})$ is a polynomial of degree L/N in these variables, and it is not a zero polynomial because setting $\mathbf{V} = \overline{\mathbf{V}}_k$ yields the valuation $P_k = \det(\mathbf{I}_{L/N}) = 1$. It follows that $P \triangleq \prod_{k\in[K]} P_k$ is a non-zero polynomial with degree KL/N. By Schwartz-Zippel Lemma, the probability of P evaluating to zero is not more than $\frac{KL/N}{q^N} = \frac{K\min_{k\in[K]} \mathrm{rk}(\mathbf{H}_k)}{q^N} < 1$. Therefore, there exists a realization of $\mathbf{V}$ for which the evaluation of P is non-zero $\implies \overline{\mathbf{U}}_k\mathbf{H}_k^T\mathbf{V}$ is invertible for all $k \in [K]$ for this realization of $\mathbf{V}$. Now let $\mathbf{U}_k \triangleq (\overline{\mathbf{U}}_k\mathbf{H}_k^T\mathbf{V})^{-1}\overline{\mathbf{U}}_k$. We obtain that $\mathbf{U}_k\mathbf{H}_k^T\mathbf{V} = \mathbf{I}_{L/N}$ for all $k \in [K]$. Now, let the input at the transmitter be $\widetilde{X} = \mathbf{V}\mathbf{W}$. Receiver $k \in [K]$ then hears $\widetilde{Y}_k = \mathbf{H}_k^T\widetilde{X} = \mathbf{H}_k^T\mathbf{V}\mathbf{W}$. The decoding at Receiver k is then $\mathbf{U}_k\widetilde{Y}_k = \mathbf{U}_k\mathbf{H}_k^T\mathbf{V}\mathbf{W} = \mathbf{W}$, which is L symbols (considered in $\mathbb{F}_q$) of the data stream. Thus, the scheme achieves $L/N = \min_k \mathrm{rk}(\mathbf{H}_k)$.

H Proof of Lemma 4

The proof is by construction. We make use of the Generalized Reed Solomon (GRS) code. Let $\mathbb{F}_q$ be a field. $n, k \in \mathbb{N}$ such that $k \leq n$. $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n) \in \mathbb{F}_q^n$, $\boldsymbol{u} = (u_1, \cdots, u_n) \in \mathbb{F}_q^n$, such that $\alpha_i \neq \alpha_j$ for $i \neq j$ and $u_i \neq 0$ for $i \in [n]$. This requires that $q \geq n$. Define

$$\mathrm{GRS}_{k,n}^q(\boldsymbol{\alpha}, \boldsymbol{u}) \triangleq \begin{bmatrix} u_1 & u_2 & u_3 & \cdots & u_n \\ u_1\alpha_1 & u_2\alpha_2 & u_3\alpha_3 & \cdots & u_n\alpha_n \\ u_1\alpha_1^2 & u_2\alpha_2^2 & u_3\alpha_3^2 & \cdots & u_n\alpha_n^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ u_1\alpha_1^{k-1} & u_2\alpha_2^{k-1} & u_3\alpha_3^{k-1} & \cdots & u_n\alpha_n^{k-1} \end{bmatrix} \in \mathbb{F}_q^{k\times n} \tag{94}$$

as the generator matrix of an $[n, k]$ GRS code over $\mathbb{F}_q$. GRS codes have the following properties [55].

1. GRS codes are MDS. Any k columns of the matrix $\text{GRS}_{k,n}^q(\boldsymbol{\alpha}, \boldsymbol{u})$ are $\mathbb{F}_q$ linearly independent.

2. The dual code of a GRS code is also a GRS code. In particular, there exists $\boldsymbol{v} = (v_1, v_2, \cdots, v_n) \in \mathbb{F}_q^n$, $v_i \neq 0$ for $i \in [n]$ such that

$$\text{GRS}_{k,n}^q(\boldsymbol{\alpha}, \boldsymbol{u}) \cdot \text{GRS}_{n-k,n}^q(\boldsymbol{\alpha}, \boldsymbol{v})^T = \mathbf{0}_{k\times(n-k)}. \tag{95}$$

Note that $\lceil N/2 \rceil + \lfloor N/2 \rfloor = N$. Define

$$\mathbf{M} = \begin{bmatrix} \text{GRS}_{\lceil N/2 \rceil,N}^q(\boldsymbol{\alpha}, \boldsymbol{u}) & \mathbf{0}_{\lceil N/2 \rceil \times N} \\ \mathbf{0}_{\lfloor N/2 \rfloor \times N} & \text{GRS}_{\lfloor N/2 \rfloor,N}^q(\boldsymbol{\alpha}, \boldsymbol{v}) \end{bmatrix} \in \mathbb{F}_q^{N\times 2N}. \tag{96}$$

We claim that this $\mathbf{M}$ is half-MDS and it is a valid transfer matrix of an N-sum box. Note that the idea of placing the generator matrices of two codes that are dual to each other on the diagonal to construct a SSO matrix follows the CSS construction [34, 35]. Now we have,

$$\begin{aligned}
&(\mathbf{M}\mathbf{J}_{2N})\mathbf{M}^T \\
&= \begin{bmatrix} \mathbf{0}_{\lceil N/2 \rceil \times N} & -\text{GRS}_{\lceil N/2 \rceil,N}^q(\boldsymbol{\alpha}, \boldsymbol{u}) \\ \text{GRS}_{\lfloor N/2 \rfloor,N}^q(\boldsymbol{\alpha}, \boldsymbol{v}) & \mathbf{0}_{\lfloor N/2 \rfloor \times N} \end{bmatrix} \begin{bmatrix} \text{GRS}_{\lceil N/2 \rceil,N}^q(\boldsymbol{\alpha}, \boldsymbol{u}) & \mathbf{0}_{\lceil N/2 \rceil \times N} \\ \mathbf{0}_{\lfloor N/2 \rfloor \times N} & \text{GRS}_{\lfloor N/2 \rfloor,N}^q(\boldsymbol{\alpha}, \boldsymbol{v}) \end{bmatrix}^T && (97) \\
&= \begin{bmatrix} \mathbf{0}_{\lceil N/2 \rceil \times \lceil N/2 \rceil} & -\text{GRS}_{\lceil N/2 \rceil,N}^q(\boldsymbol{\alpha}, \boldsymbol{u}) \cdot \text{GRS}_{\lfloor N/2 \rfloor,N}^q(\boldsymbol{\alpha}, \boldsymbol{v})^T \\ \text{GRS}_{\lfloor N/2 \rfloor,N}^q(\boldsymbol{\alpha}, \boldsymbol{v}) \cdot \text{GRS}_{\lceil N/2 \rceil,N}^q(\boldsymbol{\alpha}, \boldsymbol{u})^T & \mathbf{0}_{\lfloor N/2 \rfloor \times \lfloor N/2 \rfloor} \end{bmatrix} && (98) \\
&= \mathbf{0}_{N\times N} && (99)
\end{aligned}$$

Finally, since GRS codes are MDS, it follows that the $\mathbf{M}$ constructed in (96) is half-MDS. Therefore, if the field size $q \geq N$, there exists an N-sum box operating in $\mathbb{F}_q$ that has a half-MDS transfer matrix. □

References

[1] Y. Yao and S. A. Jafar, "The capacity of classical summation over a quantum MAC with arbitrarily replicated inputs," *CPCC Technical Report*, 2023. [Online]. Available: http://escholarship.org/uc/item/4tp227hz

[2] A. S. Cacciapuoti, M. Caleffi, R. V. Meter, and L. Hanzo, "When entanglement meets classical communications: Quantum teleportation for the quantum internet," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3808–3833, March 2020.

[3] A. S. Cacciapuoti, M. Caleffi, F. Tafuri, F. S. Cataliotti, S. Gherardini, and G. Bianchi, "Quantum internet: Networking challenges in distributed quantum computing," *IEEE Network*, vol. 34, no. 1, pp. 137–143, January/February 2020.

[4] M. Caleffi, M. Amoretti, D. Ferrari, D. Cuomo, J. Illiano, A. Manzalini, and A. S. Cacciapuoti, "Distributed quantum computing: a survey," 2022. [Online]. Available: https://arxiv.org/abs/2212.10609

[5] M. Walter, D. Gross, and J. Eisert, "Multipartite entanglement," *Quantum Information: From Foundations to Quantum Technology Applications*, pp. 293–330, 2016.

[6] J. Eisert and H. J. Briegel, "Schmidt measure as a tool for quantifying multiparticle entanglement," *Physical Review A*, vol. 64, no. 2, p. 022306, 2001.

[7] H. J. Briegel and R. Raussendorf, "Persistent entanglement in arrays of interacting particles," *Phys. Rev. Lett.*, vol. 86, pp. 910–913, Jan 2001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.86.910

[8] C. H. Bennett, D. P. DiVincenzo, J. A. Smolin, and W. K. Wootters, "Mixed-state entanglement and quantum error correction," *Phys. Rev. A*, vol. 54, pp. 3824–3851, Nov 1996. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.54.3824

[9] V. Coffman, J. Kundu, and W. K. Wootters, "Distributed entanglement," *Phys. Rev. A*, vol. 61, p. 052306, Apr 2000. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.61.052306

[10] W. K. Wootters, "Entanglement of formation of an arbitrary state of two qubits," *Phys. Rev. Lett.*, vol. 80, pp. 2245–2248, Mar 1998. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.80.2245

[11] G. Vidal, "Entanglement monotones," *Journal of Modern Optics*, vol. 47, no. 2-3, pp. 355–376, 2000. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/09500340008244048

[12] M. A. Nielsen, "Conditions for a class of entanglement transformations," *Phys. Rev. Lett.*, vol. 83, pp. 436–439, Jul 1999. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.83.436

[13] A. Shimony, "Degree of entanglement," *Annals of the New York Academy of Sciences*, vol. 755, no. 1, pp. 675–679, 1995.

[14] H. Barnum and N. Linden, "Monotones and invariants for multi-particle quantum states," *Journal of Physics A: Mathematical and General*, vol. 34, no. 35, p. 6787, 2001.

[15] T.-C. Wei and P. M. Goldbart, "Geometric measure of entanglement and applications to bipartite and multipartite quantum states," *Phys. Rev. A*, vol. 68, p. 042307, Oct 2003. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.68.042307

[16] W. Helwig, W. Cui, J. I. Latorre, A. Riera, and H.-K. Lo, "Absolute maximal entanglement and quantum secret sharing," *Physical Review A*, vol. 86, no. 5, p. 052335, 2012.

[17] M. Huber and J. I. De Vicente, "Structure of multidimensional entanglement in multipartite systems," *Physical review letters*, vol. 110, no. 3, p. 030501, 2013.

[18] D. Bruß, G. M. D'Ariano, M. Lewenstein, C. Macchiavello, A. Sen, U. Sen *et al.*, "Distributed quantum dense coding," *Physical review letters*, vol. 93, no. 21, p. 210501, 2004.

[19] N. Shutty, M. Wootters, and P. Hayden, "Tight limits on nonlocality from nontrivial communication complexity; a.k.a. reliable computation with asymmetric gate noise," in *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, nov 2020. [Online]. Available: https://doi.org/10.1109%2Ffocs46700.2020.00028

[20] D. Gavinsky, "Quantum versus classical simultaneity in communication complexity," *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6466–6483, 2019.

[21] H. Buhrman, R. Cleve, and A. Wigderson, "Quantum vs. classical communication and computation," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 63–68.

[22] G. Brassard, R. Cleve, and A. Tapp, "Cost of exactly simulating quantum entanglement with classical communication," *Physical Review Letters*, vol. 83, no. 9, p. 1874, 1999.

[23] A. Kawachi and H. Nishimura, "Communication complexity of private simultaneous quantum messages protocols," *IACR Cryptology ePrint https://eprint.iacr.org/2021/636.pdf*, May 2021.

[24] R. B. Christensen and P. Popovski, "Private product computation using quantum entanglement," *arXiv preprint arXiv:2305.05993*, 2023.

[25] A. Kawachi and H. Nishimura, "Communication complexity of private simultaneous quantum messages protocols," *arXiv preprint arXiv:2105.07120*, 2021.

[26] A. C.-C. Yao, "Quantum circuit complexity," in *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*. IEEE, 1993, pp. 352–361.

[27] D. Deutsch and R. Jozsa, "Rapid solution of problems by quantum computation," *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, vol. 439, no. 1907, pp. 553–558, 1992.

[28] B. K. Rai and B. K. Dey, "On network coding for sum-networks," *IEEE Transactions on Information Theory*, vol. 58, no. 1, pp. 50–63, 2012.

[29] A. Ramamoorthy and M. Langberg, "Communicating the sum of sources over a network," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 4, pp. 655–665, 2013.

[30] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6463–6486, 2011.

[31] R. Appuswamy, M. Franceschetti, N. Karamchandani, and K. Zeger, "Network coding for computing: Cut-set bounds," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 1015–1030, Feb. 2011.

[32] R. Appuswamy and M. Franceschetti, "Computing linear functions by linear coding over networks," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 422–431, Jan. 2014.

[33] D. Gottesman, "Stabilizer codes and quantum error correction," 1997, PhD thesis, California Institute of Technology.

[34] A. R. Calderbank and P. W. Shor, "Good quantum error-correcting codes exist," *Physical Review A*, vol. 54, no. 2, p. 1098, 1996.

[35] A. Steane, "Multiple-particle interference and quantum error correction," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 452, no. 1954, pp. 2551–2577, 1996.

[36] M. Allaix, Y. Lu, Y. Yao, T. Pllaha, C. Hollanti, and S. Jafar, "N-sum box: An abstraction for linear computation over many-to-one quantum networks," 2023. [Online]. Available: https://arxiv.org/abs/2304.07561

[37] S. Song and M. Hayashi, "Capacity of quantum private information retrieval with multiple servers," *IEEE Transactions on Information Theory*, vol. 67, no. 1, pp. 452–463, 2020.

[38] ——, "Capacity of quantum private information retrieval with colluding servers," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5491–5508, 2021.

[39] M. Allaix, S. Song, L. Holzbaur, T. Pllaha, M. Hayashi, and C. Hollanti, "On the capacity of quantum private information retrieval from MDS-coded and colluding servers," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 3, pp. 885–898, 2022.

[40] S. Song and M. Hayashi, "Capacity of quantum symmetric private information retrieval with collusion of all but one of servers," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 380–390, 2021.

[41] A. Aytekin, M. Nomeir, S. Vithana, and S. Ulukus, "Quantum symmetric private information retrieval with secure storage and eavesdroppers," 2023.

[42] N. Raviv, I. Tamo, and E. Yaakobi, "Private information retrieval in graph-based replication systems," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3590–3602, 2019.

[43] Z. Jia and S. Jafar, "On the asymptotic capacity of X-secure T-private information retrieval with graph-based replicated storage," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6280–6296, October 2020.

[44] P. Fei, Z. Chen, Z. Wang, and S. A. Jafar, "Communication-efficient clock synchronization," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 2592–2597.

[45] C. H. Bennett and S. J. Wiesner, "Communication via one- and two-particle operators on einstein-podolsky-rosen states," *Phys. Rev. Lett.*, vol. 69, pp. 2881–2884, Nov 1992. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.69.2881

[46] S. Bose, V. Vedral, and P. L. Knight, "Multiparticle generalization of entanglement swapping," *Phys. Rev. A*, vol. 57, pp. 822–829, Feb 1998. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.57.822

[47] Z. Shadman, H. Kampermann, D. Bruß, and C. Macchiavello, "Distributed superdense coding over noisy channels," *Phys. Rev. A*, vol. 85, p. 052306, May 2012. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.85.052306

[48] X. S. Liu, G. L. Long, D. M. Tong, and F. Li, "General scheme for superdense coding between multiparties," *Phys. Rev. A*, vol. 65, p. 022304, Jan 2002. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.65.022304

[49] A. S. De and U. Sen, "Quantum advantage in communication networks," *arXiv: Quantum Physics*, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:118971962

[50] M. A. Sohail, T. A. Atif, and S. S. Pradhan, "Unified approach for computing sum of sources over CQ-MAC," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 1868–1873.

[51] M. A. Sohail, T. A. Atif, A. Padakandla, and S. S. Pradhan, "Computing sum of sources over a classical-quantum MAC," *IEEE Transactions on Information Theory*, vol. 68, no. 12, pp. 7913–7934, 2022.

[52] M. Hayashi and Á. Vázquez-Castro, "Computation-aided classical-quantum multiple access to boost network communication speeds," *Physical Review Applied*, vol. 16, no. 5, p. 054021, 2021.

[53] S. de Bone, R. Ouyang, K. Goodenough, and D. Elkouss, "Protocols for creating and distilling multipartite GHZ states with bell pairs," *IEEE Transactions on Quantum Engineering*, vol. 1, pp. 1–10, 01 2020.

[54] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf, "Quantum fingerprinting," *Phys. Rev. Lett.*, vol. 87, p. 167902, Sep 2001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.87.167902

[55] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*. Elsevier, 1977, vol. 16.

[56] A. S. Holevo, "Bounds for the quantity of information transmitted by a quantum communication channel," *Problemy Peredachi Informatsii*, vol. 9, no. 3, pp. 3–11, 1973.

[57] S. Massar, S. Pironio, and D. Pitalúa-García, "Hyperdense coding and superadditivity of classical capacities in hypersphere theories," *New Journal of Physics*, vol. 17, no. 11, p. 113002, 2015.

[58] M. Pawłowski, T. Paterek, D. Kaszlikowski, V. Scarani, A. Winter, and M. Żukowski, "Information causality as a physical principle," *Nature*, vol. 461, no. 7267, pp. 1101–1104, 2009.

[59] A. S. Holevo, "The capacity of the quantum channel with general signal states," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 269–273, 1998.

[60] S.-Y. Li, R. Yeung, and N. Cai, "Linear network coding," *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 371–381, 2003.

[61] R. Ahlswede, N. Cai, S.-Y. Li, and R. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[62] M. Bahramgiri and S. Beigi, "Graph states under the action of local clifford group in non-binary case," *arXiv preprint quant-ph/0610267*, 2006.

[63] H. Sun and S. Jafar, "On the capacity of computation broadcast," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3417–3434, Jun. 2020.

[64] Y. Yao and S. A. Jafar, "On the generic capacity of K-user symmetric linear computation broadcast," 2022. [Online]. Available: https://arxiv.org/abs/2209.07602

[65] M. Hayashi and K. Wang, "Dense coding with locality restriction on decoders: Quantum encoders versus superquantum encoders," *PRX Quantum*, vol. 3, no. 3, p. 032201, Sep 2022.

[66] Y. Lu, Y. Yao, and S. A. Jafar, "On the capacity of secure K-user product computation over a quantum mac," *arXiv preprint arXiv:2305.20073*, 2023.

[67] J. Briët, H. Buhrman, M. Laurent, T. Piovesan, and G. Scarpa, "Entanglement-assisted zero-error source-channel coding," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1124–1138, 2014.

[68] H. Witsenhausen, "The zero-error side information problem and chromatic numbers (corresp.)," *IEEE Transactions on Information Theory*, vol. 22, no. 5, pp. 592–593, 1976.

[69] N. Alon, "The shannon capacity of a union," *Combinatorica*, vol. 18, no. 3, pp. 301–310, 1998.

[70] Y. Hardy and W.-H. Steeb, *Matrix Calculus, Kronecker Product and Tensor Product: A Practical Approach to Linear Algebra, Multilinear Algebra and Tensor Calculus with Software Implementations*. World Scientific, 2019.