Strong Convereses Are Just Edge Removal Properties

Oliver Kosut, Member, IEEE and Jörg Kliewer, Senior Member, IEEE

Abstract

This paper explores the relationship between two ideas in network information theory: edge removal and strong converses. Edge removal properties state that if an edge of small capacity is removed from a network, the capacity region does not change too much. Strong converses state that, for rates outside the capacity region, the probability of error converges to 1. Various notions of edge removal and strong converse are defined, depending on how edge capacity and residual error probability scale with blocklength, and relations between them are proved. In particular, each class of strong converse implies a specific class of edge removal. The opposite directions are proved for deterministic networks. Furthermore, a technique based on a novel causal version of the blowing-up lemma is used to prove that for discrete memoryless stationary networks, the weak edge removal property—that the capacity region changes continuously as the capacity of an edge vanishes—is equivalent to the exponentially strong converse—that outside the capacity region, the probability of error goes to 1 exponentially fast. This result is used to prove exponentially strong converses for several examples—including the cut-set bound and the discrete 2-user interference channel with strong interference—with only a small variation from traditional weak converse proofs.

I. INTRODUCTION

Consider a general network communication scenario given an arbitrary collection of sources and sinks connected via an arbitrary network channel. The sources are independent and each

O. Kosut is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287 USA (email: okosut@asu.edu).
J. Kliewer is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (email: jkliewer@njit.edu).
This work was presented in part at the 2016 IEEE International Symposium on Information Theory.
This material is based upon work supported by the National Science Foundation under Grant No. CCF-1439465, CCF-1440014, CNS-1526547, CCF-1453718.
source is demanded by a subset of sinks, where this subset can be different for each sink. A general interest in network information theory is to determine the capacity of such networks, defined by the set of achievable rates for each source. As this problem is known to be challenging, we consider the simpler problem of how the capacity of these networks change if only a single edge is removed from the network. This problem has first been studied by [1], [2]. The authors have shown that for acyclic noiseless networks and a variety of demand types for which the cut-set bound is tight, removing an edge of capacity $\delta$ the capacity of every min-cut is reduced by at most $\delta$ in each dimension. Further, in [3] it has been shown for a noiseless multiple multicast demand that this edge removal property also holds for generalized network sharing [4] and linear programming [5] outer bounds. In addition, the existence of the edge removal property has for example been tied to the problem whether a network coding instance allows a reconstruction with $\epsilon$ or zero error [6], [7], respectively. Another example is the connection of edge removal to the equivalency between a network coding instance and a corresponding index coding problem [8]. Recently, it has been shown that for a multiple-access channel with a so called “cooperation facilitator” [9]–[13] the edge removal property does not hold. In particular, for this setting the authors show the surprising result that adding a small capacity edge can lead to a significant increase in network capacity. These results have also been extended to networks with state [14] and to edges which can carry only a single bit over all times under the maximal error criterion [15]. However, despite significant progress has been made to understand the scenarios in which the edge removal property holds, the solution to the general problem is open.

In this work, we address the connection of edge removal to the existence of strong converses, a connection that has been explored only briefly in [16, Chap. 3]. The strong converse theorem states that the error probability converges to 1 for large blocklengths $n$ if the rate exceeds the capacity. This is in contrast to a weak converse which only indicates that the error probability is bounded away from zero if we operate at a rate beyond capacity. The benefit of strong converses are that they strengthen the interpretation of capacity as a sharp phase transition. They also allow the following interesting interpretation: if a strong converse exists for a given network instance, $\epsilon$ reliable codes (i.e., codes which allow reconstruction with $\epsilon$ error) must have rate tuples within the capacity region for $\epsilon \in [0, 1)$ and large $n$. Thus, strong converses constitute “one-and-a-half-th order” results, in that they are a step toward establishing how achievable rates change as the probability of error and blocklength scale. Strong converses have been established for numerous problems, including point-to-point settings, e.g., for discrete memoryless channels and quantum
channels. Recently it has been shown that a strong converse holds for a discrete memoryless networks with tight cut-set bounds [17].

In the following, we categorize the notions of edge removal and strong converses into different classes depending on how edge capacity and residual error probability, resp., scale with blocklength, and demonstrate relations between these instances. See Fig. 1 for a summary of our results. In particular, our contributions are as follows:

1) We show that each specific class of strong converse always implies a specific class of edge removal. This implication holds in great generality: whether the network channel model is deterministic or probabilistic, discrete or continuous, or even whether it has memory.

2) We show that implications in the opposite directions hold in some cases. In particular, we show that each opposite direction holds for deterministic networks.

3) We further show that for all discrete memoryless stationary networks, the weak edge removal property is equivalent to the exponentially strong converse. An exponentially strong converse means that for any rate vector outside the (almost) zero error capacity region, the error probability approaches 1 exponentially fast, whereas the weak edge removal property states that if remove a small edge with rate growing sublinear in the blocklength, the zero error capacity region does not change. The proof is based on a novel causal version of the blowing-up lemma [18].

4) We demonstrate that for networks composed of independent point-to-point links, a similar equivalence holds for weaker conditions—between the ordinary strong converse and what we call the very weak edge removal property, wherein the edge carries an unbounded number of bits that grows very slowly with blocklength.

5) These results, particularly the equivalence between weak edge removal and the exponentially strong converse, enable us to, without much effort, strengthen many existing computable outer bounds or weak converses to prove that they hold in an exponentially strong sense. We demonstrate this for the cut-set bound, strengthening the result of [17] to show that for rates outside the region defined by cut-set bound, the probability of error converges to 1 exponentially fast. We also prove exponentially strong converses for discrete broadcast channels, and for the discrete 2-user interference channel with strong interference.

All the above mentioned reduction results between edge removal and strong converses reveal
the surprising fact that for many cases, satisfying edge removal—a condition related only to first-order capacity—implies a seemingly stronger “one-and-a-half order” property, namely the existence of a specific version of a strong converse indicated by the leftward arrows in Fig. 1. This highlights again the power of the edge removal property.

This paper is organized as follows. We first introduce the model and definitions of various strong converse and edge removal properties in Sec. II. After that, in Sec. III we prove that strong converses imply edge removal properties. The opposite directions for deterministic networks is then proven in Sec. IV. Then, in Sec. V we prove one of the main results in this paper, namely equivalence between weak edge removal and the exponentially strong converse for discrete stationary memoryless. We then show equivalence between weak edge removal and the ordinary strong converse for networks of independent point-to-point links in Sec. VI. After that, in Sec. VII we derive several applications of our results, including the cut-set bound, broadcast channels, and interference channel. Finally, Sec. VIII offers the conclusions.

II. Model and Definitions

Notation: For an integer $k$ we define $[1 : k] = \{1, \ldots, k\}$. All logarithms and exponentials have base 2. For sequences $a_n, b_n$, we write $a_n \asymp b_n$ if $\log(a_n)/n$ and $\log(b_n)/n$ have the same limit as $n \to \infty$. Given two probability measures $P$ and $Q$ on the same alphabet $\mathcal{X}$, the relative entropy is given by

$$D(P||Q) = \int dP \log \frac{dP}{dQ}$$

(1)

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative. The total variational distance is given by

$$d_{TV}(P, Q) = \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)|.$$  

(2)

The Hamming distance between two sequences $x^n, y^n \in \mathcal{X}^n$ is denoted

$$d_H(x^n, y^n) = |\{t \in [1 : n] : x_t \neq y_t\}|.$$  

(3)

A. Network Model

We begin with a very general network model. Many of our results apply only for discrete memoryless networks or deterministic networks, but some basic results apply in much more generality.
Consider a network consisting of \(d\) nodes, where node \(i \in [1 : d]\) wishes to convey a message \(W_i\) at rate \(R_i\) to a set of destination nodes \(D_i \subseteq [1 : d]\). The channel model consists of:

- An input alphabet \(\mathcal{X}_i\) for each \(i \in [1 : d]\),
- An output alphabet \(\mathcal{Y}_i\) for each \(i \in [1 : d]\),
- For each time step \(t\), a conditional probability measure given by
  \[
  p(y_{1t}, \ldots, y_{dt} | y_{1t-1}, \ldots, y_{dt-1}, x_{1t}, \ldots, x_{dt}).
  \] (4)

**Definition 1:** A network is memoryless and stationary if the probability measure in (4) can be written
  \[
  p(y_{1t}, \ldots, y_{dt} | x_{1t}, \ldots, x_{dt})
  \] (5)
and these distributions are the same for all \(t\).

**Definition 2:** A network is deterministic if the probability measure in (4) is always deterministic.

**Definition 3:** A network is discrete if all input and output alphabets are finite sets.

For any \(R = (R_1, \ldots, R_d) \in \mathbb{R}^d\), an \((R, n)\) code consists of

- for each node \(i \in [1 : d]\) and time \(t \in [1 : n]\), an encoding function
  \[
  \phi_{it} : [1 : 2^{nR_i}] \times \mathcal{Y}_{1t-1} \rightarrow \mathcal{X}_i
  \] (6)
- for each \(i, j \in [1 : d]\) where \(j \in D_i\), a decoding function
  \[
  \psi_{ij} : [1 : 2^{nR_j}] \times \mathcal{Y}_{nt} \rightarrow [1 : 2^{nR_i}].
  \] (7)

Assume messages \(W_i\) for \(i = 1, \ldots, d\) are independent and each uniformly distributed over \([1 : 2^{nR_i}]\). The channel input from node \(i\) at time \(t\) is given by \(X_{it} = \phi_{it}(W_i, Y_{it-1})\). For \(j \in D_i\), the estimate of \(W_i\) at node \(j\) is given by \(\hat{W}_{ij} = \psi_{ij}(W_j, Y_{jn})\). We write \(W\) for the complete vector of messages, and \(\hat{W}\) for the complete vector of message estimates. Given a \((R, n)\) code, the average probability of error is
  \[
  P_e^{(n)} = \mathbb{P}(\hat{W} \neq W)
  \] (8)
where \(\hat{W} \neq W\) denotes the event that \(\hat{W}_{ij} \neq W_j\) for some \(i \in [1 : d]\), \(j \in D_i\). For blocklength \(n\) and \(\epsilon \in [0, 1]\), let \(\mathcal{R}(n)(\mathcal{N}, \epsilon) \subseteq \mathbb{R}^d\) be the set of rates \(R\) for which there exists an \((R, n)\) code

\(^1\)We assume for simplicity that at most one message originates at each node; all results can be easily generalized to the scenario in which multiple messages originate at each node.
with probability of error at most $\varepsilon$. For a sequence $\varepsilon_n \in [0, 1]$, we say a rate vector $R$ is achievable with respect to $\varepsilon_n$ if there exists an integer $n_0$ such that for all $n' \geq n_0$, $R \in \mathcal{R}^{(n')}(\mathcal{N}, \varepsilon_n')$.

The capacity region $\mathcal{R}(\mathcal{N}, \varepsilon_n)$ is given by the closure of the set of all achievable rate vectors with respect to $\varepsilon_n$. Note that $\mathcal{R}^{(n)}(\mathcal{N}, \varepsilon)$ is defined as a function of the single value $\varepsilon$, whereas $\mathcal{R}(\mathcal{N}, \varepsilon_n)$ is a function of the infinite sequence $\varepsilon_n$.

In principle $\mathcal{R}(\mathcal{N}, \varepsilon_n)$ is defined for any sequence $\varepsilon_n \in [0, 1]$. However, it will be useful to restrict ourselves to sequences for which $-\frac{1}{n} \log(1 - \varepsilon_n)$ has a limit; the following proposition, proved in Appendix A, proves that we may do this without loss of generality.

**Proposition 1:** Let $\mathcal{N}$ be any memoryless stationary network. For any $\delta > 0$, let $\varepsilon_n$ and $\tilde{\varepsilon}_n$ be two sequences where

$$\delta = \lim \inf_{n \to \infty} -\frac{1}{n} \log(1 - \varepsilon_n) = \lim \inf_{n \to \infty} -\frac{1}{n} \log(1 - \tilde{\varepsilon}_n).$$

Then $\mathcal{R}(\mathcal{N}, \varepsilon_n) = \mathcal{R}(\mathcal{N}, \tilde{\varepsilon}_n)$.

As consequence of Proposition 1 for any sequence $\varepsilon_n$ where $\delta = \lim \inf_{n \to \infty} -\frac{1}{n} \log(1 - \varepsilon_n) > 0$, $\mathcal{R}(\mathcal{N}, \varepsilon_n) = \mathcal{R}(\mathcal{N}, 1 - \exp\{-n \delta\})$. Thus, it is enough to focus on sequences $\varepsilon_n$ where either $\varepsilon_n = 1 - \exp\{-n \delta\}$ for some $\delta > 0$, or $-\log(1 - \varepsilon_n) \in o(n)$. Note that the latter includes any sequence converging to a constant in $[0, 1]$.

For fixed $\varepsilon$, $\mathcal{R}(\mathcal{N}, \varepsilon)$ denotes the capacity region with asymptotic error probability $\varepsilon$. With some abuse of notation, define the usual asymptotically-zero-error capacity region as

$$\mathcal{R}(\mathcal{N}, 0^+) = \bigcap_{\varepsilon > 0} \mathcal{R}(\mathcal{N}, \varepsilon).$$

Equivalently we may write

$$\mathcal{R}(\mathcal{N}, 0^+) = \bigcup_{\varepsilon_n \in o(1)} \mathcal{R}(\mathcal{N}, \varepsilon_n).$$

**Remark 1:** Using average probability of error rather than maximal probability of error in our definition of capacity region is not merely convenient; it is critical to many of our results. Indeed, it is illustrated in [13], [15] that edge removal characteristics are very different with maximal probability of error rather than average, and thus the relationship between edge removal and strong converses in the maximal probability of error context is likely to be different.

We proceed to define 7 different properties: 3 notions of a strong converse and 4 notions of the edge removal property. The relationships that we will prove among these properties are shown in Fig. 1.
Fig. 1. Diagram showing the relationships between various strong converses and edge removal properties. Solid black lines represent implications that always hold (Proposition 3 and Theorem 4). All the dashed or dotted lines hold for deterministic networks (Theorem 6) but do not hold in general. The red dotted line does not hold even for noisy memoryless stationary networks (Remark 2). The black dash-dotted line holds for discrete memoryless stationary networks (Theorem 9). The blue dashed line holds for discrete memoryless stationary networks made up of independent point-to-point links (Theorem 13), and we conjecture that it holds for all discrete memoryless stationary networks.

B. Strong Converses

We define three different versions of the strong converse, beginning with the usual strong converse.

**Definition 4:** Network $\mathcal{N}$ satisfies the strong converse if for all $R \notin \mathcal{R}(\mathcal{N},0^+)$, any sequence of $(R,n)$ codes has probability of error approaching 1 as $n \to \infty$. Equivalently, $\mathcal{R}(\mathcal{N},\epsilon) = \mathcal{R}(\mathcal{N},0^+)$ for all $\epsilon \in (0,1)$.

The following definition originates from [16, Chap. 3].

**Definition 5:** Network $\mathcal{N}$ satisfies the exponentially strong converse if for all $R \notin \mathcal{R}(\mathcal{N},0^+)$, any sequence of $(R,n)$ codes has probability of error approaching 1 exponentially fast. Equivalently, for any sequence $\epsilon_n \in (0,1)$ for which $-\log(1-\epsilon_n) \in o(n)$, $\mathcal{R}(\mathcal{N},\epsilon_n) = \mathcal{R}(\mathcal{N},0^+)$. 
Finally, we define the following even stronger condition.

**Definition 6:** For any \( R \not\in R(N, 0^+) \), define the distance to the capacity region \( \beta(R) \) as the smallest number \( \beta \) such that \( R - \beta 1 \in R(N, 0^+) \). Network \( N \) satisfies the extremely strong converse if there exists a finite constant \( K \) depending only on the network \( N \) such that, for any \( R \not\in R(N, 0^+) \), any sequence of \((R, n)\) codes has probability of error satisfying \( \epsilon_n \geq 1 - 2^{-n \beta(R)/K} \). Equivalently, for any \( \epsilon_n \) where \( 1 - \epsilon_n = 2^{-n \alpha} \), if \( R \in R(N, \epsilon_n) \) then \( R - K \alpha 1 \in R(N, 0^+) \).

It is easy to see that the extremely strong converse implies the exponentially strong converse, which in turn implies the ordinary strong converse.

**Remark 2:** Exponential bounds on the probability of success for rates above capacity for point-to-point channels were first considered in [19]. Later, [20] exactly characterized the optimal exponent of the success probability for rates above capacity. Namely, [20] showed that for a discrete-memoryless point-to-point channel \( p(y|x) \) with capacity \( C \), for all \( R > C \) the optimal probability of error \( \epsilon_n \) satisfies
\[
1 - \epsilon_n = 2^{-\alpha(R)n}
\]
where
\[
\alpha(R) = \min_{q(x)} \min_{v(y|x)} D(v(y|x)||p(y|x)|q(x)) + |R - I_{q\times v}(X;Y)|^+
\]
(12)
where \( I_{q\times v}(X;Y) \) is the mutual information between \( X \) and \( Y \) where \( (X,Y) \sim q(x)v(y|x) \), and \( |\cdot|^+ \) represents the positive part. Intuitively, \( v(y|x) \) represents an empirical conditional distribution; decoding will be possible if the channel behaves like one with capacity greater than \( R \) (i.e. when the second term in (12) is zero), and the first term in (12) is the exponential rate of the probability that channel \( p(y|x) \) behaves like \( v(y|x) \) with input distribution \( q \).

This result constitutes an exponentially strong converse in our terminology, since \( \alpha(R) > 0 \) for all \( R > C \), but interestingly it is not an extremely strong converse for most noisy channels. Note that an extremely strong converse is equivalent to \( \alpha'(C) > 0 \), where \( \alpha' \) is the first derivative of the function \( \alpha \). However, as we show in the following proposition (proved in Appendix B) this holds if and only if \( C = \log \min \{|X|, |Y|\} \).

**Proposition 2:** For a discrete-memoryless point-to-point channel with capacity \( C \), if \( C < \log \min \{|X|, |Y|\} \) then \( \alpha'(C) = 0 \); if \( C = \log \min \{|X|, |Y|\} \) then \( \alpha(R) = R - C \), i.e., \( \alpha'(C) > 0 \).

Note that channels where \( C = \log \min \{|X|, |Y|\} \) are essentially noiseless (i.e., bit-pipes), so noisy channels do not satisfy the extremely strong converse. Thus, while we are able to show equivalence between the extremely strong converse and the strong edge removal property for
Fig. 2. The modified network for edge removal properties. Nodes $a$ and $b$ are connected to every other link in the network by infinite capacity links, while the link between them is limited to only $k_n$ bits. Edge removal properties hold when the capacity region of this network is unchanged when the link between $a$ and $b$ is removed.

deterministic networks (see Fig. 1), this equivalence cannot hold for most noisy networks, as the extremely strong converse simply does not hold.

C. Edge Removal Properties

For a subset of nodes $\mathcal{V} \subseteq [1 : d]$ and an integer sequence $k_n$, we define a modified network $\mathcal{N}(\mathcal{V}, k_n)$, illustrated in Fig. 2, as follows: Start with $\mathcal{N}$, and add two nodes denoted $a$ and $b$. For each node $i \in \mathcal{V}$, add an infinite capacity link from $i$ to $a$, and an infinite capacity link from $b$ to $i$. Finally, add a bit-pipe from $a$ to $b$ that can noiselessly transmit $k_n$ bits total, where $n$ is the blocklength of the code and $k_n/n$ need not be constant. In the case that $k_n$ is not an integer multiple of $n$, this bit-pipe cannot be modeled as a stationary memoryless channel. Instead, we assume that the $k_n$ bits are scheduled such that after $t$ timesteps, $\left\lfloor \frac{k_n}{n} t \right\rfloor$ have been transmitted; that is, at time $t$, the link is allowed to transmit exactly

$$\left\lfloor \frac{k_n}{n} t \right\rfloor - \left\lfloor \frac{k_n}{n} (t - 1) \right\rfloor$$

(13)

These are special nodes in that messages do not originate at them. Thus the capacity region of $\mathcal{N}(\mathcal{V}, k_n)$ has the same dimension as that of $\mathcal{N}$. 
Let $\mathcal{R}_V(\mathcal{N}, \epsilon_n, k_n) = \mathcal{R}(\mathcal{N}(V, k_n), \epsilon_n)$. For the most part we are interested in the case that $V = [1 : d]$, we define for convenience $\mathcal{R}(\mathcal{N}, \epsilon_n, k_n) = \mathcal{R}(\mathcal{N}(V, k_n), \epsilon_n)$. For any $k_n$, it is certainly true that $\mathcal{R}(\mathcal{N}, \epsilon_n) \subseteq \mathcal{R}(\mathcal{N}, \epsilon_n, k_n)$. Note also that $\mathcal{R}(\mathcal{N}, \epsilon_n, 0) = \mathcal{R}(\mathcal{N}, \epsilon_n)$.

Roughly, edge removal properties state that for small $k_n$, the capacity of network $\mathcal{N}(V, k_n)$ is not too different from that of $\mathcal{N}$. To be precise, we define four different versions of this property, beginning with the weakest.

**Definition 7:** Network $\mathcal{N}$ satisfies the extremely weak edge removal property if, for any constant integer $k$, $\mathcal{R}(\mathcal{N}, 0^+, k) = \mathcal{R}(\mathcal{N}, 0^+)$.  

**Definition 8:** Network $\mathcal{N}$ satisfies the very weak edge removal property if $\bigcap_{k_n: k_n \to \infty} \mathcal{R}(\mathcal{N}, 0^+, k_n) = \mathcal{R}(\mathcal{N}, 0^+)$.  

**Definition 9:** Network $\mathcal{N}$ satisfies the weak edge removal property if, for any $k_n \in o(n)$, $\mathcal{R}(\mathcal{N}, 0^+, k_n) = \mathcal{R}(\mathcal{N}, 0^+)$.  

**Definition 10:** Network $\mathcal{N}$ satisfies the strong edge removal property if there exists a constant $K$ depending only on the network $\mathcal{N}$ such that for all $\delta > 0$, if $R \in \mathcal{R}(\mathcal{N}, 0^+, \delta n)$, then $R - K\delta \in \mathcal{R}(\mathcal{N}, 0^+)$.  

**Remark 3:** When $V = [1 : d]$, then node $a$ has complete knowledge of every signal sent in the network, so the link $(a, b)$ can be used to simulate any other small-capacity link. In particular, for any network $\mathcal{N}'$ consisting of $\mathcal{N}$ supplemented by a link (or multiple links) with total capacity at most $k_n$ bits, then $\mathcal{R}(\mathcal{N}', \epsilon_n) \subseteq \mathcal{R}(\mathcal{N}, \epsilon_n, k_n)$. One example of such a network $\mathcal{N}'$ is one that allows for rate-limited feedback. For this reason, one consequence of edge removal results are outer bounds on networks with rate-limited feedback.

**Remark 4:** The extremely weak edge removal property, wherein the extra edge carries a bounded number of bits as the blocklength grows, appears in none of our results proving relationships to strong converses. Nevertheless, we have chosen to include this definition because it is a natural one, and indeed the property seems tantalizingly likely to be true for all realistic systems. However, it was shown in [15] that for maximal error probability, there exists a network where the extremely weak property does not hold. This again points to the contrast between average and maximal error probability. In light of our other results, the extremely weak property

---

3 One could imagine other models, such as where the bit transmission schedule is flexible but chosen in advance by the code, or where the schedule can be chosen at run-time. These model variations are unlikely to impact results, but here we adopt the more restrictive model.
also presents an interesting question: namely, is it equivalent to some version of a strong converse? Based on our results that for some networks, the very weak edge removal property is equivalent to the ordinary strong converse, if there is an equivalent converse to the extremely weak property, it appears that it would need to be weaker than the ordinary strong converse, but perhaps stronger than the ordinary weak converse. No such property has occurred to us.

The following proposition, proved in Appendix C, proves some simple facts about the edge removal properties, including justifying the names of the edge removal properties.

**Proposition 3:**

1) The strong edge removal property implies the weak edge removal property, which implies the very weak edge removal property, which implies the extremely weak edge removal property.

2) For any \( V \subseteq [1 : d] \),

\[
\bigcap_{k_n : k_n \to \infty} \mathcal{R}_V (N, 0^+, k_n) = \bigcap_{\epsilon > 0} \bigcup_{k \in \mathbb{N}} \mathcal{R}_V (N, \epsilon, k). \tag{15}
\]

3) For any \( V \subseteq [1 : d] \),

\[
\bigcup_{k_n \in o(n)} \mathcal{R}_V (N, 0^+, k_n) = \bigcap_{\delta > 0} \mathcal{R}_V (N, 0^+, \delta n). \tag{16}
\]

### III. Deriving Edge Removal Properties from Strong Converse

The following theorem states that each of the three strong converse properties implies one of the edge removal properties. This result holds for our most general network model.

**Theorem 4:** For any network \( N \), the following hold:

1) The strong converse implies very weak edge removal.
2) The exponentially strong converse implies weak edge removal.
3) The extremely strong converse implies strong edge removal.

Statement (2) of this theorem was proved for noiseless networks in [16, Sec. 3.3]. Our proof is a generalized version of theirs. We begin with the following lemma.

**Lemma 5:** For any integers \( n \) and \( k \) and any \( \epsilon \in [0, 1] \),

\[
\mathcal{R}^{(n)} (N, \epsilon, k) \subseteq \mathcal{R}^{(n)} (N, 1 - (1 - \epsilon)2^{-k}). \tag{17}
\]

**Proof:** Let \( R \in \mathcal{R}^{(n)} (N, \epsilon, k) \), so there is an \( n \)-length code with rate vector \( R \) and probability of error at most \( \epsilon \) on network \( N([1 : d], k) \). We convert this code to one on network \( N \) as follows.
Under the code on $\mathcal{N}([1 : d], k)$, let $X_{ab}$ be the message sent on the link from node $a$ to node $b$. Recall that $X_{ab} \in \{0, 1\}^k$. Let $E$ be the overall error event for network $\mathcal{N}([1 : d], k)$. We have

$$1 - \epsilon \leq P(E^c) = \sum_{x_{ab} \in \{0, 1\}^k} P(X_{ab} = x_{ab})P(E^c | X_{ab} = x_{ab}).$$

(18)

There must be some $x_{ab}^* \in \{0, 1\}^k$ for which

$$P(X_{ab} = x_{ab}^*)P(E^c | X_{ab} = x_{ab}^*) \geq (1 - \epsilon)2^{-k}.$$  

(19)

Construct a code for network $\mathcal{N}$ that behaves exactly like the original code on network $\mathcal{N}([1 : d], k)$, except that all nodes assume that node $b$ received the signal $x_{ab}^*$. Let $P_e$ be the probability of error for this code. Note that with probability $P(X_{ab} = x_{ab}^*)$, the code’s behavior will be just as if the code on $\mathcal{N}([1 : d], k)$ were in effect. Thus

$$1 - P_e \geq P(X_{ab} = x_{ab}^*)P(E^c | X_{ab} = x_{ab}^*) \geq (1 - \epsilon)2^{-k}.$$ 

(20)

Therefore $R \in R^{(n)}(\mathcal{N}, 1 - (1 - \epsilon)2^{-k})$.

**Proof of Theorem 4:** We first show statement (1). Assume the strong converse holds. Thus

$$\bigcap_{\epsilon > 0} \bigcup_{k \in \mathbb{N}} R(\mathcal{N}, \epsilon, k) \subseteq \bigcap_{\epsilon \in (0, 1)} \bigcup_{k \in \mathbb{N}} R(\mathcal{N}, 1 - (1 - \epsilon)2^{-k})$$

(21)

$$= \bigcap_{\epsilon > 0} \bigcup_{k \in \mathbb{N}} R(\mathcal{N}, 0^+)$$

(22)

$$= R(\mathcal{N}, 0^+)$$

(23)

where (21) follows from Lemma 5; (22) follows from the strong converse, because $1 - (1 - \epsilon)2^{-k} \in (0, 1)$ for any $\epsilon \in (0, 1)$ and $k \in \mathbb{N}$; and (23) follows because $R(\mathcal{N}, 0^+)$ is closed. Therefore, very weak edge removal holds by statement 2 of Proposition 3.

We now prove statement (2). Assume the exponentially strong converse holds. For any $k_n \in o(n)$, we have

$$R(\mathcal{N}, 0^+, k_n) = \bigcap_{\epsilon > 0} R(\mathcal{N}, \epsilon, k_n)$$

$$\subseteq \bigcap_{\epsilon > 0} R(\mathcal{N}, 1 - (1 - \epsilon)2^{-k_n})$$

(24)

$$\subseteq \bigcup_{\epsilon_n : -\log(1 - \epsilon_n) \in o(n)} R(\mathcal{N}, \epsilon_n)$$

(25)

$$\subseteq R(\mathcal{N}, 0^+)$$

(26)
where (24) follows from Lemma 5, (25) from the fact that $k_n \in o(n)$, and (26) from the exponentially strong converse. Therefore weak edge removal holds.

We now prove statement (3). Assume the extremely strong converse holds. For any $\delta > 0$ we have

$$R(N, 0^+, \delta n) = \bigcap_{\epsilon > 0} R(N, \epsilon, \delta n) \subseteq \bigcap_{\epsilon > 0} R(N, 1 - (1 - \epsilon)2^{-\delta n})$$

(27)

where (27) follows from Lemma 5. Note that $(1 - \epsilon)2^{-\delta n} \equiv 2^{-\delta n}$. Thus if $R \in R(N, 0^+, \delta n)$, then, by the extremely strong converse, $R - K\delta \in R(N, 0^+)$ for some constant $K$. Therefore strong edge removal holds.

IV. Deterministic Networks

The following is our main theorem for deterministic networks, asserting that each edge removal property is equivalent to a strong converse property.

Theorem 6: For any deterministic network $N$, the following hold:

1) The very weak edge removal property holds if and only if the strong converse holds.

2) The weak edge removal property holds if and only if the exponentially strong converse holds.

3) The strong edge removal property holds if and only if the extremely strong converse holds.

To prove Theorem 6 we begin with several lemmas. The following lemma is a simple consequence of Markov’s inequality, but will be instrumental in proving that edge removal properties imply strong converses.

Lemma 7: Let $X$ be a real-valued random variable where $X \leq x_{\max}$ a.s. If $\mathbb{E}X \geq \mu \geq 0$, then

$$P\left(X \geq \frac{\mu}{2}\right) \geq \frac{\mu}{2x_{\max}}.$$  

(28)

Proof: We have

$$P\left(X \geq \frac{\mu}{2}\right) = 1 - P\left(X < \frac{\mu}{2}\right)$$

(29)

$$= 1 - P\left(x_{\max} - X > x_{\max} - \frac{\mu}{2}\right)$$

(30)

$$\geq 1 - \frac{x_{\max} - \mathbb{E}X}{x_{\max} - \mu/2}$$

(31)

$$\geq \frac{\mu/2}{x_{\max} - \mu/2}$$

(32)
\[ \geq \frac{\mu}{2x_{\max}} \]  

where (31) follows from Markov’s inequality, (32) since \( \mathbb{E}X \leq \mu \), and (33) since \( \mu \geq 0 \).

The following lemma provides the core result that is needed to prove Theorem 6. The proof is adapted from that of [21, Lemma 2].

**Lemma 8:** Let \( \mathcal{N} \) be a deterministic network. There exists a finite-valued function \( \eta(\tilde{\epsilon}) \) for any \( \tilde{\epsilon} \in (0, 1) \) and a constant \( K \) such that, for any sequence \( \epsilon \in [0, 1) \), any \( n \), and any \( \tilde{\epsilon} \in (0, 1) \),

\[ \mathcal{R}^{(n)}(\mathcal{N}, \epsilon) \subseteq \mathcal{R}^{(n)}(\mathcal{N}, \tilde{\epsilon}, \eta(\tilde{\epsilon}) - K \log(1 - \epsilon)). \]  

**Proof:** Let \( \mathbf{R} \in \mathcal{R}^{(n)}(\mathcal{N}, \epsilon) \). Fix \( \tilde{\epsilon} \) and define with hindsight

\[ k = \left\lceil d + \log \ln \frac{4d}{\tilde{\epsilon}} - \log(1 - \epsilon) \right\rceil. \]  

To prove the lemma, it is enough to show that for sufficiently large \( n \),

\[ \mathbf{R} \in \mathcal{R}^{(n)}(\mathcal{N}, \tilde{\epsilon}, 3dk), \]  

because \( 3dk \leq \eta(\tilde{\epsilon}) - K \log(1 - \epsilon) \) where \( \eta(\tilde{\epsilon}) = 3d(d + 1) + 3d \log \ln \frac{4d}{\tilde{\epsilon}} \) and \( K = 3d \).

Define a rate vector \( \tilde{\mathbf{R}} = (\tilde{R}_1, \ldots, \tilde{R}_d) \) where

\[ \tilde{R}_i = \begin{cases} 
R_i - \frac{k}{n}, & R_i \geq \frac{2k}{n} \\
0, & R_i < \frac{2k}{n}.
\end{cases} \]  

We will prove that

\[ \tilde{\mathbf{R}} \in \mathcal{R}^{(n)}(\mathcal{N}, \tilde{\epsilon}, dk) \]  

by constructing a code of rate \( \tilde{\mathbf{R}} \) on network \( \mathcal{N}([1 : d], dk) \). This is enough to prove (36), since \( nR_i - n\tilde{R}_i \leq 2k \), so we may simply expand the edge from node \( a \) to \( b \) to carry \( 2dk \) additional bits, adding up to \( 2k \) bits for each message.

Since \( \mathbf{R} \in \mathcal{R}^{(n)}(\mathcal{N}, \epsilon) \), there exists a code of length \( n \) at rate vector \( \mathbf{R} \) and probability of error at most \( \epsilon \). For \( i = 1, \ldots, d \), let \( \mathcal{W}_i = [2^{nR_i}] \) be the message set for the \( i \)th message, and let

\[ \mathcal{W} = \prod_{i=1}^d \mathcal{W}_i \]  

be the set of complete message vectors \( \mathbf{w} = (w_1, \ldots, w_d) \). Let \( R = \sum R_i \), so \( |\mathcal{W}| = 2^n R \). Since the network is deterministic, whether or not an error occurs depends entirely on the message vector \( \mathbf{w} \in \mathcal{W} \) that is chosen. Let \( \Gamma \) be the subset of \( \mathcal{W} \) of message vectors that do not lead
to errors. Thus the probability of error is precisely $1 - 2^{-nR|\Gamma|}$. By the assumption that the probability of error is at most $\epsilon$, we have that

$$|\Gamma| \geq |\mathcal{W}|(1 - \epsilon) = 2^{nR}(1 - \epsilon). \quad \text{(40)}$$

Recall that $\tilde{R}_i = 0$ if $nR_i < 2k$, so this message is not significant. For ease of notation, we assume for now that $nR_i \geq 2k$ for all messages $i$, so that $\tilde{R}_i = R_i - \frac{k}{n}$. We employ a random binning argument. For each $i$, randomly and uniformly choose a partition of $\mathcal{W}_i$

$$P_i(1), \ldots, P_i(2^{n\tilde{R}_i}), \quad \text{(41)}$$

from among all partitions for which $|P_i(\tilde{w}_i)| = 2^k$ for all $\tilde{w}_i \in [1 : 2^{n\tilde{R}_i}]$. Given these partitions, the code proceeds as follows. Given messages $\tilde{W}_1, \ldots, \tilde{W}_d$, they are all transmitted to node $a$. Node $a$ then looks for an element of

$$\Gamma \cap \prod_{i=1}^{d} P_i(\tilde{W}_i). \quad \text{(42)}$$

If there is no such element, declare an error. Otherwise, let $\mathcal{W} = (W_1, \ldots, W_d)$ be one such element. For each $i$, let $I_i \in \{1, \ldots, 2^k\}$ be the index of $W_i$ in the set $P_i(\tilde{W}_i)$. Node $a$ determines $I_i$ for each $i$ and transmits $(I_1, \ldots, I_d)$ to node $b$. Note that the number of bits required is $dk$.

At the originating source node for message $i$, $W_i$ can be determined from $\tilde{W}_i$ and $I_i$. Subsequently, the code proceeds as if $\mathcal{W}$ were the true message vector. When a destination node $j$ produces a message estimate $\hat{W}_{ij}$, it constructs the final message estimate as the $\hat{\tilde{W}}_{ij} \in [1 : 2^{n\tilde{R}_i}]$ such that $\hat{W}_{ij} \in P_i(\hat{W}_{ij})$. Since by assumption $\mathcal{W} \in \Gamma$, there is no error as long as (42) is not empty.

For $\tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_d)$ let

$$q(\tilde{w}) \triangleq \mathbb{P}\left(\Gamma \cap \prod_{i=1}^{d} P_i(\tilde{w}_i) = \emptyset\right) \quad \text{(43)}$$

where the probability is with respect to the random choice of partitions $P_i$. We proceed to show that $q(\tilde{w}) \leq \bar{\epsilon}$ for all $\tilde{w}$. This proves that there exists at least one code with probability of error $\bar{\epsilon}$.

For each $i \in [1 : d]$, define for all $w_1, \ldots, w_{i-1}$, the set

$$A_i(w_1, \ldots, w_{i-1}) = \left\{w_i : \left|\{ (w_{i+1}, \ldots, w_d) : (w_1, \ldots, w_d) \in \Gamma \} \right| \geq (1 - \epsilon)2^{n(R_{i+1} + \cdots + R_d) - i}\right\}. \quad \text{(44)}$$
Fix \( w_1, \ldots, w_{i-1} \) such that \( w_{i-1} \in A_{i-1}(w_1, \ldots, w_{i-2}) \). Define the random variable

\[
X(w_1, \ldots, w_{i-1}) = |\{(w_{i+1}, \ldots, w_d) : (w_1, \ldots, w_{i-1}, W_i, w_{i+1}, \ldots, w_d) \in \Gamma\}|. \tag{45}
\]

where as usual \( W_i \) is uniformly distributed on \([1 : 2^{|R_i}|]\). Note that

\[
\mathbb{E}X(w_1, \ldots, w_{i-1}) = 2^{-nR_i} \sum_{w_i} |\{(w_{i+1}, \ldots, w_d) : (w_1, \ldots, w_d) \in \Gamma\}| \tag{46}
\]

\[
= 2^{-nR_i} |\{(w_1, \ldots, w_d) : (w_1, \ldots, w_d) \in \Gamma\}| \tag{47}
\]

\[
\geq (1 - \epsilon)2^{n(R_{i+1} + \cdots + R_d) - (i-1)} \tag{48}
\]

where the inequality follows from the assumption that \( w_{i-1} \in A_{i-1}(w_1, \ldots, w_{i-2}) \). Hence

\[
|A_i(w_1, \ldots, w_{i-1})| = 2^{nR_i} \mathbb{P}\left(X(w_1, \ldots, w_{i-1}) \geq (1 - \epsilon)2^{n(R_{i+1} + \cdots + R_d) - i}\right) \tag{49}
\]

\[
\geq (1 - \epsilon)2^{nR_i-i} \tag{50}
\]

where (50) follows from Lemma 7 and the fact that \( X \leq 2^{n(R_{i+1} + \cdots + R_d)} \).

Fix \( \tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_d) \). For each \( i = 1, \ldots, d \), define

\[
Q_i = \{(w_1, \ldots, w_i) : w_j \in P_j(\tilde{w}_j) \cap A_j(w_1, \ldots, w_{j-1}) \text{ for all } j \leq i\}. \tag{51}
\]

Note that \( Q_d \subseteq \Gamma \cap \prod_{i=1}^d P_i(\tilde{w}_i) \), so

\[
q(\tilde{w}) \leq \mathbb{P}(Q_d = \emptyset) \leq \sum_{i=1}^d \mathbb{P}(Q_i = \emptyset | Q_{i-1} \neq \emptyset). \tag{52}
\]

If \( Q_{i-1} \neq \emptyset \), then there exists \( (w_1, \ldots, w_{i-1}) \in Q_{i-1} \), and the probability that \( P_i(\tilde{w}_i) \cap A_i(w_1, \ldots, w_{i-1}) = \emptyset \) is given by

\[
\frac{\binom{2^{nR_i} - |A_i(w_1, \ldots, w_{i-1})|}{2^k}}{2^{nR_i}} \leq \frac{(2^{nR_i} - (1 - \epsilon)2^{nR_i-i})^2}{2^{nR_i}} \tag{53}
\]

where we have applied the bound (50). Thus

\[
q(\tilde{w}) \leq \sum_{i=1}^d \frac{2^{nR_i-(1-\epsilon)2^{nR_i-i}}}{2^k} \tag{54}
\]

\[
= \sum_{i=1}^d \frac{(2^{nR_i} - (1 - \epsilon)2^{nR_i-i})!(2^{nR_i} - 2^k)!}{(2^{nR_i} - 2^{nR_i-i} - 2^k)!)(2^{nR_i})!} \tag{55}
\]

\[
\leq \sum_{i=1}^d \frac{(2^{nR_i} - (1 - \epsilon)2^{nR_i-i})2^k}{(2^{nR_i} - 2^k)^2} \tag{56}
\]

\[
= \sum_{i=1}^d \exp \left\{2^k \log(1 - (1 - \epsilon)2^{-i}) - 2^k \log(1 - 2^{k-nR_i})\right\} \tag{57}
\]
\[ \leq \sum_{i=1}^{d} \exp \left\{ -2^{k-d}(1 - \epsilon) \log e - 2^k \log(1 - 2^{k-nR_i}) \right\} \quad (58) \]

\[ \leq \sum_{i=1}^{d} \frac{\tilde{\epsilon}}{4d} \exp \left\{ -2^k \log(1 - 2^{-k}) \right\} \quad (59) \]

\[ \leq \sum_{i=1}^{d} \frac{\tilde{\epsilon}}{4d} \exp \left\{ -2^k \log(1 - 2^{-k}) \right\} \quad (60) \]

\[ \leq \tilde{\epsilon} \quad (61) \]

where (58) follows since \( \log(1 + x) \leq x \log e \), (59) follows from the choice of \( k \) in (35), (60) follows by the assumption that \( R_i \geq \frac{2k}{n} \) for all \( i \), and (61) follows since \( \exp \left\{ -2^k \log(1 - 2^{-k}) \right\} \leq 4 \) for any \( k \geq 1 \).

Proof of Theorem 6: Theorem 4 proves that each strong converse property implies the corresponding edge removal property, so we only need to prove the opposite directions.

Suppose the very weak edge removal property holds. For any constant \( \epsilon \), applying Lemma 8 gives

\[ \mathcal{R}(\mathcal{N}, \epsilon) \subseteq \bigcap_{\tilde{\epsilon} > 0} \mathcal{R}(\mathcal{N}, \tilde{\epsilon}, \eta(\tilde{\epsilon}) - K \log(1 - \epsilon)) \quad (62) \]

\[ \subseteq \bigcap_{\tilde{\epsilon} > 0} \bigcup_{k \in \mathbb{N}} \mathcal{R}(\mathcal{N}, \tilde{\epsilon}, k). \quad (63) \]

\[ = \mathcal{R}(\mathcal{N}, 0^+) \quad (64) \]

where the last equality holds by very weak edge removal. Therefore the strong converse holds.

Now suppose the weak edge removal property holds. For any sequence \( \epsilon_n \) where \(- \log(1 - \epsilon_n) \in o(n)\), applying Lemma 8 gives

\[ \mathcal{R}(\mathcal{N}, \epsilon_n) \subseteq \bigcap_{\tilde{\epsilon} > 0} \mathcal{R}(\mathcal{N}, \tilde{\epsilon}, \eta(\tilde{\epsilon}) - K \log(1 - \epsilon_n)) \quad (65) \]

\[ \subseteq \bigcap_{\tilde{\epsilon} > 0} \mathcal{R}(\mathcal{N}, \tilde{\epsilon}, \sqrt{n} - K \log(1 - \epsilon_n)) \quad (66) \]

\[ = \mathcal{R}(\mathcal{N}, 0^+, \sqrt{n} - K \log(1 - \epsilon_n)) \quad (67) \]

\[ = \mathcal{R}(\mathcal{N}, 0^+) \quad (68) \]

where the last equality follows from weak edge removal, since \( \sqrt{n} - K \log(1 - \epsilon_n) \in o(n) \). Therefore the exponentially strong converse holds.
Finally, suppose the strong edge removal property holds. For any $\alpha > 0$, let $\epsilon_n$ where $1 - \epsilon_n = 2^{-n\alpha}$. Applying Lemma 8 gives

$$R(N, \epsilon_n) = R(N, 1 - 2^{-n\alpha})$$

$$\subseteq \bigcap_{\tilde{\epsilon} > 0} R(N, \tilde{\epsilon}, \eta(\tilde{\epsilon}) + K\alpha n)$$

$$\subseteq \bigcap_{\tilde{\epsilon} > 0} R(N, \tilde{\epsilon}, (K + 1)\delta n)$$

$$= R(N, 0^+, (K + 1)\delta n).$$

where (69) follows from Prop. 11 and (70) follows from Lemma 8. Thus, if $R \in R(N, \epsilon_n)$, by strong edge removal $R - K'(K + 1)\alpha n \in R(N, 0^+)$. Therefore the extremely strong converse holds.

V. Discrete Stationary Memoryless Networks

The following is our main theorem for discrete stationary memoryless networks, connecting the exponentially strong converse to the weak edge removal property. In addition, we show that both these properties are equivalent to an even weaker form of the weak edge removal property—namely, where the nodes $a$ and $b$ connect only to transmitting nodes; i.e. those nodes $i$ where $X_i \neq \emptyset$. This is a generalization of the “cooperation facilitator” model from [9]–[14], which connected only to the transmitters in a multiple-access channel, but not the receiver. The intuition behind connecting only to transmitting nodes is that the extra edge is only useful for nodes when encoding. At a node that decodes but does not transmit, a decoder achieving a probability of error $\epsilon_n$ that does not approach 1 exponentially fast can be converted into a list decoder with small probability of error and a relatively small list. Thus, a short hash can be transmitted to identify the correct message and thereby reduce the probability of error to zero with a unique decoder. This approach cannot be used at nodes that must transmit, which is why the extra edge can in principle be helpful.

Theorem 9: For any discrete stationary memoryless network $N$, the following three statements are equivalent:

1) The exponentially strong converse holds.

2) The weak edge removal property holds.

3) For any $k_n \in o(n)$, $R_V(N, 0^+, k_n) = R(N, 0^+)$, where $V$ is the set of nodes $i$ such that $X_i \neq \emptyset$. 

Observe that statement 1 of the theorem implies statement 2 by Theorem 4. Statement 2 implies statement 3 by the definition of the weak edge removal property. Hence it remains only to show that statement 3 implies statement 1. The main tool in doing so will be a modified version of the blowing-up lemma. The blowing-up lemma, originally proved in [22] (see also [18], [23]), has been used in the proof of numerous strong converse results. In some sense our result is a generalization of this technique. The traditional blowing-up lemma is stated as follows.

**Lemma 10:** Let \( X^n \in \mathcal{X}^n \) be a sequence of independent random variables. Fix \( A \subseteq \mathcal{X}^n \) where \( P_{X^n}(A) = \exp\{-n\gamma_n\} \) for a sequence \( \gamma_n \to 0 \). For any \( \ell \), define the **blown-up** version of \( A \) as

\[
A_\ell = \{ x^n : d_H(x^n, y^n) \leq \ell \text{ for some } y^n \in A \}
\]

where recall \( d_H \) is the Hamming distance. There exist a sequence \( \delta_n \to 0 \) where \( P_{X^n}(A_{\ell \delta_n}) \to 1 \).

The following is a causal version of the blowing-up lemma. It is stronger than the usual blowing-up lemma, but it follows almost directly from Marton’s proof of the blowing-up lemma in [18]. We provide a proof for completeness.

**Lemma 11:** Let \( X^n \in \mathcal{X}^n \) be a random sequence, not necessarily independent. Fix \( A \subseteq \mathcal{X}^n \). There exists a sequence of conditional distributions \( P_{Z_t|Y_t, Z_{t-1}} \) for \( t = 1, \ldots, n \) such that, if we let \( Y^n \in \mathcal{X}^n, Z^n \in \mathcal{X}^n \) have joint distribution

\[
P_{Y^n, Z^n}(y^n, z^n) = \prod_{t=1}^{n} P_{X_t|X_{t-1}}(y_t|z_{t-1}) P_{Z_t|Y_t, Z_{t-1}}(z_t|y_t, z_{t-1})
\]

then \( Z^n \in A \) almost surely, and

\[
\mathbb{E}d_H(Y^n, Z^n) \leq \sqrt{\frac{n}{2 \log e} \log \frac{1}{P_{X^n}(A)}}.
\]

**Proof:** Let \( \tilde{X}^n \) be a random sequence with distribution that of \( X^n \) conditioned on the set \( A \). That is,

\[
P_{\tilde{X}^n}(x^n) = \begin{cases} 
P_{X^n}(x^n) & x^n \in A \\ 0 & x^n \notin A. \end{cases}
\]

For any \( t \in [1 : n] \) and \( z^{t-1} \in \mathcal{X}^{t-1} \), by [24] Theorem 1 there exists a pair of random variables \( X_t(z^{t-1}), \tilde{X}_t(z^{t-1}) \) with joint distribution \( P_{X_t(z^{t-1}), \tilde{X}_t(z^{t-1})} \) such that the marginal distributions satisfy

\[
P_{X_t(z^{t-1})} = P_{X_t|X^{t-1}=z^{t-1}},
\]
\[ P_{X_t(z^{t-1})} = P_{X_t|X^{t-1}=z^{t-1}} \] (79)

and their joint distribution satisfies

\[ \mathbb{P}(X_t(z^{t-1}) \neq \tilde{X}_t(z^{t-1})) = d_{TV}(P_{X_t|X^{t-1}=z^{t-1}}, P_{X_t|\tilde{X}^{t-1}=z^{t-1}}). \] (80)

We now define

\[ P_{Z_t|Y_t,Z^{t-1}}(z_t|y_t,z^{t-1}) = P_{\tilde{X}_t(z^{t-1})|X_t(z^{t-1})}(z_t|y_t). \] (81)

Let \( Y^n, Z^n \) have distribution given by (75). Note that

\[ P_{Z_t|Z^{t-1}}(z_t|z^{t-1}) = \sum_{y_t} P_{X_t|X^{t-1}}(y_t|z^{t-1})P_{Z_t|Y_t,Z^{t-1}}(z_t|y_t,z^{t-1}) \] (82)

\[ = \sum_{y_t} P_{X_t(z^{t-1})}(y_t)P_{\tilde{X}_t(z^{t-1})|X_t(z^{t-1})}(z_t|y_t) \] (83)

\[ = P_{\tilde{X}_t(z^{t-1})}(z_t) \] (84)

\[ = P_{X_t,\tilde{X}^{t-1}}(z_t|z^{t-1}). \] (85)

Thus \( Z^n \) and \( \tilde{X}^n \) have the same distribution. In particular, \( Z^n \in A \) almost surely. We now have

\[ \mathbb{E}d_H(Y^n, Z^n) = \sum_{t=1}^n \mathbb{P}(Y_t \neq Z_t) \] (86)

\[ = \sum_{t=1}^n \sum_{z^{t-1}} P_{Z^{t-1}}(z^{t-1}) \sum_{y_t \neq z_t} P_{X_t|X^{t-1}}(y_t|z^{t-1})P_{Z_t|Y_t,Z^{t-1}}(z_t|y_t,z^{t-1}) \] (87)

\[ = \sum_{t=1}^n \sum_{z^{t-1}} P_{Z^{t-1}}(z^{t-1}) \sum_{y_t \neq z_t} P_{X_t(z^{t-1}),\tilde{X}_t(z^{t-1})}(y_t, z_t) \] (88)

\[ = \sum_{t=1}^n \sum_{z^{t-1}} P_{Z^{t-1}}(z^{t-1})\mathbb{P}(X_t(z^{t-1}) \neq \tilde{X}_t(z^{t-1})) \] (89)

\[ = \sum_{t=1}^n \sum_{z^{t-1}} P_{Z^{t-1}}(z^{t-1})d_{TV}(P_{X_t|X^{t-1}=z^{t-1}}, P_{\tilde{X}_t|\tilde{X}^{t-1}=z^{t-1}}) \] (90)

\[ \leq \sum_{t=1}^n \sum_{z^{t-1}} P_{Z^{t-1}}(z^{t-1}) \sqrt{\frac{1}{2}D(P_{X_t|X^{t-1}=z^{t-1}}\|P_{X_t|X^{t-1}=z^{t-1}})} \] (91)

\[ \leq n \sqrt{\frac{1}{2n} \sum_{t=1}^n \sum_{z^{t-1}} P_{Z^{t-1}}(z^{t-1})D(P_{\tilde{X}_t|\tilde{X}^{t-1}=z^{t-1}}\|P_{X_t|X^{t-1}=z^{t-1}})} \] (92)

\[ = \sqrt{\frac{n}{2} \sum_{t=1}^n \sum_{z^{t-1}} P_{\tilde{X}^{t-1}}(z^{t-1})D(P_{\tilde{X}_t|\tilde{X}^{t-1}=z^{t-1}}\|P_{X_t|X^{t-1}=z^{t-1}})} \] (93)
\[ = \sqrt{\frac{n}{2 \log e} D(P_{X^n} \parallel P_{X^n})} \]
\[ = \sqrt{\frac{n}{2 \log e} \log \frac{1}{P_{X^n}(A)}} \]

where (90) holds by (80), (91) holds by Pinsker’s inequality, (92) holds by concavity of the square root, (93) holds because \( Z^n \) and \( \tilde{X}^n \) have the same distribution, (94) holds by the chain rule for relative entropy, and (95) holds because
\[
\frac{P_{\tilde{X}^n}(\tilde{X}^n)}{P_{X^n}(X^n)} = \frac{1}{P_{X^n}(A)} \quad \text{a.s.} \tag{96}
\]

**Remark 5:** Lemma 10 can be derived from Lemma 11 as follows. If in Lemma 11, \( X^n \) is a sequence of independent random variables, then by (75), \( Y^n \) has the same distribution as \( X^n \). Thus
\[
P_{X^n}(A_\ell) = P_{Y^n}(A_\ell) \tag{97}
\]
\[
\geq \mathbb{P}(d_H(Y^n, Z^n) \leq \ell) \tag{98}
\]
\[
\geq 1 - \frac{1}{\ell} \mathbb{E}d_H(Y^n, Z^n) \tag{99}
\]
\[
\geq 1 - \frac{1}{\ell} \sqrt{\frac{n}{2 \log e} \log \frac{1}{P_{X^n}(A)}} \tag{100}
\]

where (98) holds because \( Z^n \in A \) almost surely, (99) holds by Markov’s inequality, and in (100) we have applied (76). Assuming \( P_{X^n}(A) = \exp\{-n \gamma_n\} \) where \( \gamma_n \to 0 \), if we choose, for example, \( \delta_n = \gamma_n^{1/4} \), we have \( \delta_n \to 0 \)
\[
P_{X^n}(A_{n\delta_n}) \geq 1 - \frac{\gamma_n^{1/4}}{\sqrt{2 \log e}} \to 1. \tag{101}
\]

This proves Lemma 10.

With Lemma 11 in hand, we complete the proof of Theorem 9 with the following lemma, which immediately implies that statement 3 of the theorem implies statement 1.

**Lemma 12:** For any discrete stationary memoryless network \( \mathcal{N} \),
\[
\bigcup_{\epsilon_n : -\log(1 - \epsilon_n) \in o(n)} \mathcal{R}(\mathcal{N}, \epsilon_n) \subseteq \bigcup_{k_n \in o(n)} \mathcal{R}_V(\mathcal{N}, 0^+, k_n). \tag{102}
\]

**Proof:** Fix some sequence \( \epsilon_n \) where \( -\log(1 - \epsilon_n) \in o(n) \). By statement 3 of Proposition 8 it is enough to show that for all \( \delta > 0 \), \( \mathcal{R}(\mathcal{N}, \epsilon_n) \subseteq \mathcal{R}_V(\mathcal{N}, 0^+, \delta n) \), where again \( V \) is the set of nodes \( i \) such that \( \mathcal{X}_i \neq \emptyset \). Let \( R \) be achievable with respect to \( \epsilon_n \). Thus for sufficiently large
There exists an \( n \)-length code with average probability of error at most \( \epsilon_n \). Let \((\phi_{it}, \psi_{ij})\) be the encoding/decoding functions for this code (see (6)–(7)). We describe a new code, illustrated in Fig. 3 achieving the same rate vector with vanishing probability of error on the network \( \mathcal{N}(\mathcal{V}, n\delta) \). Note that for any \( i \in \mathcal{V}^c \), we have \( X_i = \emptyset \), so if \( R_{i} > 0 \) the probability of success would be exponentially small; thus we must have \( R_{i} = 0 \).

Network stacking: We adopt the notion of network stacking from [25]. The motivation for using network stacking is that it allows us to convert an arbitrary coding operation in a single
time instance into a coding operation across a long block, thereby taking advantage of the law of large numbers. In particular, we construct $N$ copies of the original $n$-length code, using a total of $nN$ channel uses. Each copy is referred to as a “layer”, indexed by an integer $\ell \in [1 : N]$.

Unlike a block Markov approach, in which one would transmit an $n$-length block corresponding to the original code in sequence, in the network stacking approach we transmit $N$ copies of a single time instance $t \in [1 : n]$ of the original code before moving on to the next one. Thus coding can be done “across the layers”, using the fact that the $N$ copies of any symbol are i.i.d., while maintaining the causal structure of the original code.

We use underlines to indicate symbols on the stacked network. In particular, $X_{it}(\ell)$ is the transmitted symbol from node $i$ at time $t$ in layer $\ell$; $X^n(\ell)$ refers to the $n$-length sequence of symbols in layer $\ell$; $X_{it}$ refers to the $N$-length sequence of symbols at time $t$ in all layers; $X^n$ refers to the full $nN$-length sequence of all layers and time instances. We define $Y_{it}(\ell)$, etc. similarly. Moreover, $W_{ia}(\ell)$ is the message originating at node $i$ in layer $\ell$, and $W_i$ is the complete vector of messages originating at node $i$ across all $N$ layers.

**Code phases:** Given the original $n$-length code, we construct an $N$-fold stacked code as follows, where the precise dependence between $n$ and $N$ is to be determined. The code consists of $2n + 2$ phases, each consisting of a number of timesteps: first a message coordination phase, following by $n$ transmission phases alternating with $n$ correction phases, and concluding with a hashing phase. In the message coordination phase, nodes coordinate to choose a message vector in each layer with a relatively large probability of success; this is done in exactly the same manner as for deterministic networks in Lemma 8. Each transmission phase corresponds to one timestep $t \in [1 : n]$ in the original code: the layers act independently, each performing the coding functions from the original code at time $t$. In the following correction phase, node $a$ transmits data to node $b$, describing replacements for certain received data in sub-network $V$. Node $b$ then disperses this data to the nodes in $V$; in subsequent transmission phases, nodes in $V$ use this replaced data in their coding operations. In the final hashing phase, hashes of all messages are dispersed to all nodes, in particular allowing nodes in $V^c$ to decode.

The message coordination phase consists of $O(N(-\log(1 - \epsilon_n) + \log n))$ timesteps. Each transmission phase consists of exactly $N$ timesteps, since each layer transmits exactly once. Correction phases have variable lengths, depending on how much correction data is required,
but a total of $N n \gamma_n$ timesteps are allocated for all correction phases, where

$$\gamma_n = \left( - \log \frac{1 - \epsilon_n}{4n} \right)^{1/4}. \quad (103)$$

The hashing phase consists of $O(\sqrt{\gamma_n} n N)$ timesteps. Note that in total, the transmission phases consist of $n N$ timesteps. Recalling that $- \log(1 - \epsilon_n) \in o(n)$, $\gamma_n \to 0$ as $n \to \infty$, so all other phases consist of a negligible number of timesteps.

**Message coordination phase:** For each message vector $w$, let $P_c(w)$ be the probability of correctly decoding $w$. Let

$$\Gamma = \left\{ w : P_c(w) \geq \frac{1 - \epsilon_n}{2} \right\}. \quad (104)$$

Note that

$$1 - \epsilon_n \leq \frac{1}{2nR} \sum_w P_c(w) \quad (105)$$

where $R = \sum_{i=1}^{d} R_i$. Applying Lemma 7 to the random variable $P_c(W)$, we find

$$|\Gamma| \geq \frac{1 - \epsilon_n}{2} 2^{2nR}. \quad (106)$$

In the message coordination phase, we use an identical outer code as in Lemma 8 to ensure that only message vectors in $\Gamma$ are ever used. By the same binning argument as in the proof of Lemma 8, this requires only $O(- \log(1 - \epsilon_n) + \log n)$ bits on the link $(a, b)$ for each layer. Note that nodes $a$ and $b$ are only required to contact the nodes in $\mathcal{V}$, since nodes in $\mathcal{V}^c$ have no message originating at them. We may therefore assume throughout the rest of this argument that $\mathbf{W}(\ell) \in \Gamma$ for each $\ell \in [1 : N]$.

**Correction codebook:** For each message vector $w$ of the original $n$-length code, let $Q(w)$ be the set of channel outputs $y^n_{\mathcal{V}}$ such that, for message vector $w$, $\hat{W}_{ij} = w_i$ for all $i \in [1 : d]$, $j \in \mathcal{V} \cap D_i$, and

$$\mathbb{P}(\mathbf{W} = w \mid \mathbf{W} = w, Y^n_{\mathcal{V}} = y^n_{\mathcal{V}}) \geq \frac{1 - \epsilon_n}{4}. \quad (107)$$

Since channel inputs $X^n_{\mathcal{V}}$ are deterministic functions of $Y^n_{\mathcal{V}}$, the only randomness in the probability in (107) are the channel outputs $Y^n_{\mathcal{V}^c}$ given the inputs $X^n_{\mathcal{V}}$. Recalling that $X_i = \emptyset$ for $i \in \mathcal{V}^c$, $Y^n_{\mathcal{V}^c}$ is an independent sequence given $X^n_{\mathcal{V}}$. Note that for any $w \in \Gamma$,

$$\sum_{y^n_{\mathcal{V}}} P_{Y^n_{\mathcal{V}} \mid W = w}(y^n_{\mathcal{V}}) \mathbb{P}(\mathbf{W} = w \mid \mathbf{W} = w, Y^n_{\mathcal{V}} = y^n_{\mathcal{V}}) = \mathbb{P}(\mathbf{W} = w \mid \mathbf{W} = w) = P_c(w) \quad (108)$$

$$= P_c(w) \quad (109)$$
Thus, applying Lemma 7 gives
\[ P_{Y^w} | W = w(Q(w)) \geq \frac{1 - \epsilon_n}{2}. \] (110)

We now apply Lemma 11 to the distribution \( P_{Y^w} | W = w \) and the set \( Q(w) \) to find conditional distributions \( P_{Z_{V,t} | Y_{V,t}, Z_{V,t}} \) for all \( t = [1 : n] \). Note that these distributions depend on the message vector \( w \). For each \( y_{V,t} \in Y_V \) and \( z_{V,t-1} \in Y_{V,t-1} \), independently draw

\[ f_t(w, y_{V,t}, z_{V,t-1}) \sim P_{Z_{V,t} | Y_{V,t}, Z_{V,t-1}}. \] (112)

These functions constitute a codebook known to all nodes.

**Hashing codebook:** For each \( i \in V \) and each \( w_i \in [1 : 2^{nR_i}]^N \), independently and uniformly draw \( g_i(w_i) \) from \([1 : 2^{nN \sqrt{\gamma_n}}]\). These hashing functions also constitute a codebook known to all nodes.

**Transmission phases:** Before the \( t \)th transmission phase, each node \( i \in V \) has determined \( Z_{i,t-1} \in Y_{i,t-1} \), which represent the corrected versions of its received signals (see description below of the correction phases). For each \( \ell \in [1 : N] \), node \( i \) determines and transmits

\[ X_{i,t}(\ell) = \phi_{it}(W_i(\ell), Z_{i,t-1}) \] (113)

For each \( i \in [1 : d] \), let \( Y_{i,t}(\ell) \) be the corresponding received signals.

**Correction phases:** In the correction phase after the \( t \)th transmission phase, node \( a \) learns \( Y_{i,t} \) from each \( i \in V \), and determines, for each \( \ell \in [1 : N] \),

\[ Z_{V,t}(\ell) = f_t(W(\ell), Y_{V,t}(\ell), Z_{V,t-1}(\ell)). \] (114)

For each \( \ell \) for which \( Z_{V,t}(\ell) \neq Y_{V,t}(\ell) \), node \( a \) transmits to node \( b \) a bit string with 0 followed by \( \lceil \log N | Y \rceil \) bits identifying the layer \( \ell \in [1 : N] \) as well as the value of \( Z_{V,t}(\ell) \in Y_V \). After doing this for each layer where \( Z_{V,t}(\ell) \neq Y_{V,t}(\ell) \), node \( a \) transmits the stop bit 1, signaling that all nodes should proceed to the next transmission phase. Node \( b \) then forwards this data to each node \( i \in V \). For all layers \( \ell \) for which no correcting signal was sent, each node \( i \in V \) simply sets \( Z_{i,t}(\ell) = Y_{i,t}(\ell) \).

**Hashing phase:** Node \( a \) computes \( g_i = g_i(w_i) \) for all \( i \in V \), and transmits these values to node \( b \), which subsequently disperses them to nodes in \( V \).

\(^4\)One could also compute the hash for message \( i \) directly at node \( i \), and distribute the hash to all decoder nodes from there. We choose to compute the hash at node \( a \) makes merely to make distribution of the hashes simpler to describe.
total of $d \sqrt{\gamma_n} nN$ bits, which is sub-linear in $nN$. Thus they can be transmitted over the link $(a,b)$ as long as $\delta > 0$. For each node $i \in \mathcal{V}^c$, if there exists a node $j \in \mathcal{V}$ where the point-to-point channel from $X_j$ to $Y_j$ has positive capacity, then we use a point-to-point channel code to transmit the hashes from node $j$ to node $i$. If there is no such node $j \in \mathcal{V}$, then if $i \in \mathcal{D}_k$ for any $k \in [1 : d]$, it must be that $R_k = 0$. Since the hashes occupy a sub-linear number of bits, transmitting these hashes to each node in $\mathcal{V}^c$ takes a sub-linear number of timesteps, and can be done with arbitrarily small probability of error.

**Decoding:** For each $i, j \in \mathcal{V}$ where $j \in \mathcal{D}_i$ and each $\ell \in [1 : N]$, node $j$ determines

$$W_j(\ell) = \psi_{ij}(W_j(\ell), Z^n_j(\ell)).$$

(115)

Now consider $i \in [1 : d]$ and $j \in \mathcal{V}^c \cap \mathcal{D}_i$ and each $i \in [1 : d]$ where $j \in \mathcal{D}_i$. Given $Y^n_j$ and $g_i$, find the unique $\hat{w}_i$ where $g_i = g_i(\hat{w}_i)$ and there exists $\tilde{y}_j^n$ where $\psi_{ij}(W_j(\ell), \tilde{y}_j^n(\ell)) = \tilde{w}_i(\ell)$ for each $\ell \in [1 : N]$ and

$$d_H(Y^n_j, \tilde{y}_j^n) \leq Nn\gamma_n.$$  

(116)

If there is no such $\hat{w}_i$ or more than one, declare an error.

**Probability of error analysis:** Consider the following error events

$$\mathcal{E}_1 = \{ \text{number of timesteps used in correction phases exceeds } Nn\gamma_n \}$$  

(117)

and, for $i \in [1 : d]$ and $j \in \mathcal{V}^c \cap \mathcal{D}_i$,

$$\mathcal{E}_{2ij} = \{ \psi_{ij}(W_j(\ell), \tilde{y}_j^n(\ell)) = W_i(\ell) \text{ for all } \ell \in [1 : N], \text{ for some } \tilde{y}_j^n \}$$

where $d_H(Y^n_j, \tilde{y}_j^n) \leq Nn\gamma_n$.

$$\mathcal{E}_{3ij} = \{ \psi_{ij}(W_j(\ell), \tilde{y}_j^n(\ell)) = \tilde{w}'(\ell) \text{ for all } \ell \in [1 : N], \text{ for some } \tilde{w}' \neq W_i \}$$

where $g_i(\tilde{w}') = g_i(W_i)$ and $\tilde{y}_j^n$ where $d_H(Y^n_j, \tilde{y}_j^n) \leq Nn\gamma_n$.

(118)

(119)

Note that as long as $\mathcal{E}_1$ does not occur, then by Lemma $\bigcap \mathcal{E}^{ij} \in \mathcal{Q}(\mathcal{W}(\ell))$ for all $\ell$. By the definition of $\mathcal{Q}(\mathcal{w})$, this ensures that $W_j = w_i$ for all $j \in [1 : d]$ and $i \in \mathcal{V}$. Events $\mathcal{E}_{2ij}, \mathcal{E}_{3ij}$ cover all errors that can occur at nodes in $\mathcal{V}^c$. Hence the probability of error of the overall code, averaged over random coding choices, is

$$P_e \leq \mathbb{P} \left( \bigcup_{i \in [1 : d], j \in \mathcal{V}^c \cap \mathcal{D}_i} \mathcal{E}_{2ij} \cup \mathcal{E}_{3ij} \right)$$

$$\leq \mathbb{P}(\mathcal{E}_1) + \sum_{i \in [1 : d], j \in \mathcal{V}^c \cap \mathcal{D}_i} \left[ \mathbb{P}(\mathcal{E}_{2ij} | \mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_{3ij} | \mathcal{E}_1^c) \right].$$

(120)

(121)
We first consider $E_1$. The number of bits transmitted across link $(a,b)$ during the $t$th correction phase is
\[ d_H(Y_{V,t}, Z_{V,t})(\log N|Y_V|) + 1 \]
where the final +1 accounts for the stop bit. Thus the number of bits transmitted during all $n$ correction phases is
\[ d_H(Y^n_{V}, Z^n_{V})(\log N|Y_V|) + 1 + n. \]
Recall link $(a,b)$ has capacity $\delta > 0$, meaning it can transmit a bit roughly every $1/\delta$ timesteps (cf. (13)). Thus we can bound $E_1$ by
\[ \mathbb{P}(E_1) = \mathbb{P} \left( \frac{1}{\delta} d_H(Y^n_{V}, Z^n_{V})(\log N|Y_V|) + 1 + n > Nn\gamma_n \right) \]
which vanishes since $-\gamma_n \log \gamma_n \to 0$ as $\gamma_n \to 0$.

Now we consider events $E_{2ij}, E_{3ij}$. Recall that if $E_1$ does not occur, then $Z^n_{V}(\ell) \in Q(W(\ell))$ for all $\ell$. By the definition of $Q(w)$ in (107), we have, for any $y^n_{V} \in Q(w)$
\[ \frac{1 - \epsilon_n}{4} \leq P_{w|y^n_{V}}(w|y^n_{V}, y^n_{V}) \]
\[ = \sum_{y^n_{V_c}} P_{y^n_{V_c}|y^n_{V}=y^n_{V}, w=w}(y^n_{V_c}) 1(\psi_{ij}(y^n_{V})) = w_i \text{ for all } i \in V, j \in V^c \cap D_i). \]
Note that given $Y^n_{V} = y^n_{V}$ and $W = w$, $X^n_{V}$ is determined. Since $\mathcal{X}_i = \emptyset$ for all $i \in V^c$, this conditioning also determines $X^n_{1:d}$. Thus, the distribution $P_{Y^n_{V} | X^n_{V}}$ is independent.
Applying the blowing up lemma to this distribution and the set of \( y_{\mathcal{V}^c} \) that cause all messages to be decoded correctly in \( \mathcal{V}^c \), there exists a random sequence \( Z^n_{\mathcal{V}^c} \in \mathcal{Y}^n_{\mathcal{V}^c} \) that causes all messages to be decoded correctly, and

\[
\mathbb{E}d_H(Y^n_{\mathcal{V}^c}, Z^n_{\mathcal{V}^c}) \leq \sqrt{-n \log \frac{1 - \epsilon_n}{4}} = n\gamma_n^2.
\]  

(133)

In particular, if we produce \( N \) copies of this \( Z^n_{\mathcal{V}^c} \) sequence for each layer, then Markov’s inequality gives

\[
\mathbb{P}\left(d_H(Y^n_{\mathcal{V}^c}, Z^n_{\mathcal{V}^c}) > Nn\gamma_n \right) \leq \frac{Nn\gamma_n^2}{Nn\gamma_n} = \gamma_n.
\]

(134)

In particular, for each \( i \in [1 : d] \) and \( j \in \mathcal{V}^c \cap \mathcal{D}_i \), with probability at least \( 1 - \gamma_n \), there exists \( \tilde{\nu}_j^n \) that satisfies the Hamming distance condition \( (116) \), and is decoded correctly to \( w_i \). Thus \( \mathbb{P}(E_{3ij} | E_{1i}) \) vanishes. We now consider \( E_{3ij} \). The number of messages \( w'_j \) that are considered is upper bounded by the number of sequences \( \tilde{\nu}_j^n \) satisfying \( (116) \), which is given by

\[
\sum_{k=0}^{\lfloor Nn\gamma_n \rfloor} \binom{Nn}{k} |\mathcal{Y}_i|^k \leq \exp\{nN(H(\gamma_n) + \gamma_n \log |\mathcal{Y}_i|)\}
\]

(135)

where \( H(\cdot) \) is the binary entropy function. The probability that any given \( w'_j \neq W_j \) agrees with the hash value \( g_j \) is \( 2^{-nN\sqrt{\gamma_n}} \), so

\[
\mathbb{P}(E_{3ij} | E_{1i}) \leq \exp\{nN(H(\gamma_n) + \gamma_n \log |\mathcal{Y}_i|) - nN\sqrt{\gamma_n}\}
\]

(136)

\[
\leq \exp\{-nN\sqrt{\gamma_n} / 2\}
\]

(137)

\[
= \exp\{-n\gamma^{-3/2} / 2\}
\]

(138)

where (137) holds for sufficiently large \( n \), since \( \gamma_n \to 0 \) and \( \lim_{p \to 0} H(p) / \sqrt{p} = 0 \), and (138) holds again by the choice \( N = \gamma_n^{-2} \). Since \( n\gamma^{-3/2} \to \infty \) as \( n \to \infty \), \( \mathbb{P}(E_{3ij} | E_{1i}) \) vanishes.

Remark 6: The blowing-up lemma does not appear to be strong enough to prove that the very weak edge removal property implies the ordinary strong converse. Were we to apply the same argument above to the case \( \epsilon_n = \epsilon \in (0, 1) \), in the key application of the blowing-up lemma in (126), we would have

\[
\mathbb{E}d_H(Y^n_{\mathcal{V}^c}, Z^n_{\mathcal{V}^c}) \leq \sqrt{-n \log \frac{1 - \epsilon}{2}}.
\]

(139)

This suggests that at least \( O(\sqrt{n}) \) bits per layer would be required on the extra link. However, very weak edge removal requires that we achieve the same capacity region using \( any k_n \) sequence of bits converging to infinity, even a sequence growing smaller than \( \sqrt{n} \).
VI. NETWORKS OF INDEPENDENT POINT-TO-POINT LINKS

We now consider the setting of network equivalence [25], in which $\mathcal{N}$ consists of a stationary memoryless network made up of independent point-to-point (noisy) links. Let $\mathcal{N}'$ be the same network in which each noisy point-to-point link is replaced by a noiseless bit-pipe of the same capacity. The basic result of network equivalence states that $\mathcal{R}(\mathcal{N}, 0^+) = \mathcal{R}(\mathcal{N}', 0^+)$. Theorem 9 already asserts that for such networks, the weak edge removal property holds if and only if the exponentially strong converse holds. The following theorem proves that, for such networks with acyclic topology, the same holds for the “lower level” in Fig. 1, i.e. the very weak edge removal property and the ordinary strong converse. The proof, given in Appendix 12, makes use of the network equivalence principle to connect codes on $\mathcal{N}$ to codes on $\mathcal{N}'$, and then applies Theorem 6 on $\mathcal{N}'$.

**Theorem 13:** For a discrete stationary memoryless network $\mathcal{N}$ consisting of independent point-to-point links with acyclic topology, the very weak edge removal property holds if and only if the strong converse holds.

VII. APPLICATIONS

A. Outer Bounds

Consider any outer bound $\mathcal{R}_{\text{out}}(\mathcal{N})$ for the memoryless stationary network $\mathcal{N}$; i.e. where $\mathcal{R}(\mathcal{N}, 0^+) \subseteq \mathcal{R}_{\text{out}}(\mathcal{N})$. Suppose we could show

$$\bigcup_{k_n \in o(n)} \mathcal{R}_V(\mathcal{N}, 0^+, k_n) \subseteq \mathcal{R}_{\text{out}}(\mathcal{N})$$

where as usual $V$ is the set of nodes $i$ where $\mathcal{X}_i \neq \emptyset$. In other words, the outer bound is continuous with respect to the capacity of the extra edge; that is, the outer bound satisfies a weak edge removal property. Then, applying Lemma 12, we immediately find

$$\bigcup_{\epsilon_n : -\log(1-\epsilon_n) \in o(n)} \mathcal{R}(\mathcal{N}, \epsilon_n) \subseteq \mathcal{R}_{\text{out}}(\mathcal{N}).$$

This suggests that the outer bound holds in an exponentially strong sense; that is, for any rate vector outside $\mathcal{R}_{\text{out}}(\mathcal{N})$, the probability of error approaches 1 exponentially fast.

An outer bound may also satisfy a strong edge removal property, meaning that for some constant $K$ and any $\delta$,

$$\mathcal{R}(\mathcal{N}, 0^+, \delta n) \subseteq \mathcal{R}_{\text{out}}(\mathcal{N}) + K\delta.$$
We have no equivalence between the strong edge removal property and the extremely strong converse for general noisy networks, but we do for deterministic networks. Thus, applying Lemma 8 if a deterministic network satisfies (142), then the outer bound holds in an *extremely strong* sense; that is, for any rate vector outside $\mathcal{R}_{\text{out}}(\mathcal{N})$, the probability of error approaches 1 at an exponential rate linear in the distance to the outer bound.

For many outer bounds (indeed, almost every computable outer bound that we know of), (140) can be proved without much difficulty, and in some cases the stronger statement (142) can be proved as well. This implies that most outer bounds for discrete memoryless networks hold in an exponentially strong sense, and many outer bounds for deterministic networks hold in an extremely strong sense. We illustrate this for several outer bounds (or weak converse arguments) in the next few subsections.

**B. Cut-set Bound**

Recall that the *cut-set outer bound* is given by

$$
\mathcal{R}(\mathcal{N}, 0^+) \subseteq \mathcal{R}_{\text{cut-set}}(\mathcal{N}) = \bigcup_{p(x_1, \ldots, x_d)} \left\{ \mathbf{R} : \sum_{i \in S : D_i \cap S^c \neq \emptyset} R_i \leq I(X_S; Y_{S^c} | X_{S^c}) \text{ for all } S \subseteq [1 : d] \right\}.
$$

We prove (142) for this bound. This allows us to reproduce the result of [17], that the strong converse holds for networks with tight cut-set bound. In fact, [17] proves the stronger result that the cut-set bound—even if it is not tight—holds in a *strong sense*, meaning that for any rate vector outside $\mathcal{R}_{\text{cut-set}}(\mathcal{N})$, the probability of error goes to 1. In proving (142), we conclude that for discrete memoryless networks the cut-set bound holds in an *exponentially strong* sense. Furthermore, we conclude that for deterministic networks, the cut-set bound holds in an *extremely strong* sense.

Fix some sequence $k_n$, and let $\mathbf{R} \in \mathcal{R}(0^+, k_n)$. Consider a code achieving this rate vector, and let $Z_t$ be the symbol sent along edge $(a, b)$ at time $t$, or $\emptyset$ if there is no symbol at time $t$. Note $H(Z^n) \leq k_n$. Fix any cut set $S \subseteq [1 : d]$, and let $S^c = [1 : d] \setminus S$. Also let $\mathcal{T}$ be the set of message flows that cross the cut; that is, the set of $i \in S$ where $D_i \cap S^c \neq \emptyset$. We may write

$$
\sum_{i \in \mathcal{T}} R_i = H(M_\mathcal{T})
$$

\begin{equation}
\leq I(M_\mathcal{T}; Y^n_{S^c}, Z^n) + n \epsilon_n
\end{equation}

(144)
\[ \sum_{t=1}^{n} I(M_T; Y_{S^c,t}, Z_t | Y_{S^c}^{t-1}, Z^{t-1}) + n \epsilon_n \] (146)

\[ = \sum_{t=1}^{n} I(M_T; Y_{S^c,t}, Z_t | Y_{S^c}^{t-1}, Z^{t-1}, X_{S^c,t}) + n \epsilon_n \] (147)

\[ \leq \sum_{t=1}^{n} I(M_T, Y_{S^c}^{t-1}, X_{S^c,t}; Y_{S^c,t} | Z^{t-1}, X_{S^c,t}) + n \epsilon_n \] (148)

\[ \leq \sum_{t=1}^{n} [I(M_T, Y_{S^c}^{t-1}, X_{S^c,t}; Y_{S^c,t} | Z^{t-1}, X_{S^c,t}) + H(Z_t | Z^{t-1})] + n \epsilon_n \] (149)

\[ \leq \sum_{t=1}^{n} I(X_{S^c,t}; Y_{S^c,t} | X_{S^c,t}) + H(Z^n) + n \epsilon_n \] (150)

\[ \leq n I(X_S; Y_{S^c} | X_{S^c}, Q) + k_n + n \epsilon_n \] (151)

\[ \leq n I(X_S; Y_{S^c} | X_{S^c}) + k_n + n \epsilon_n \] (152)

where (145) follows from Fano’s inequality, where \( \epsilon_n \to 0 \) as \( n \to \infty \); (147) follows since \( X_{S^c,t} \)
is a function of \( Y_{S^c}^{t-1} \) and \( Z^{t-1} \); (150) follows from the memorylessness and causality of the
network model; and (151) follows by defining \( Q \sim \text{Unif}[1 : n] \), \( X_i = X_i,Q \), and \( Y_i = Y_i,Q \), and
by the fact that \( H(Z^n) \leq k_n \). Recalling that \( \epsilon_n \to 0 \), we have

\[ \mathcal{R}_V(N, 0^+, k_n) \subseteq \mathcal{R}_{\text{cut-set}}(N) + \lim_{n \to \infty} \frac{k_n}{n}. \] (153)

In particular, (142) holds with \( K = 1 \). This in turn implies (140). Therefore, for discrete
memoryless stationary networks, the cut-set bound holds in an exponentially strong sense, and
for deterministic networks, the cut-set bound holds in an extremely strong sense.

These facts allow us to immediately derive strong converse results for various problems for
which the cut-set bound is tight. For example, based on the known tightness of the cut-set bound
for various problems:

1) the exponentially strong converse holds for reversely degraded or semideterministic relay
channels,

2) the extremely strong converse holds for finite-field deterministic multicast networks.

C. Broadcast Channel

A broadcast channel is a network where \( Y_1 = \emptyset \), \( X_i = \emptyset \) for all \( i > 1 \), and we allow multiple
messages to originate at node 1, each to be decoded at a subset of nodes in \([2 : d]\). Note that this
model includes scenarios where there are private messages, public messages, and/or messages
intended for some decoders but not all. We claim that the weak edge removal property and the exponentially strong converse hold for discrete memoryless broadcast channels. Indeed, the $\mathcal{V}$ set in Theorem 9 is simply $\{1\}$. Thus, for any sequence $k_n$ (whether or not it is in $o(n)$), $\mathcal{R}_{\{1\}}(\mathcal{N}, 0^+, k_n) = \mathcal{R}(\mathcal{N}, 0^+)$, simply because if nodes $a$ and $b$ can only communicate with node 1, then any processing done at nodes $a$ and $b$ can simply be reproduced internally at node 1 instead. Theorem 9 immediately proves the claim.

As far as we know, this is the first time a strong (or exponentially strong) converse has been proved for a problem for which the capacity region is not known in general. For the broadcast channel this is not hard to do even without the machinery developed in this paper, but it is noteworthy how easy it is to prove given Theorem 9.

D. Discrete 2-User Interference Channel with Strong Interference

A 2-user interference channel, illustrated in Fig. 4 is a network with 4 nodes, where $\mathcal{Y}_1 = \mathcal{Y}_2 = \mathcal{X}_3 = \mathcal{X}_4 = \emptyset$, $\mathcal{D}_1 = \{3\}$, and $\mathcal{D}_2 = \{4\}$. Note that, to be consistent with the notation in the rest of the paper, the received symbol by the node decoding the first message is $Y_3$, rather than $Y_1$ as it would be traditionally.

Recall that an interference channel has strong interference if

$$I(X_1; Y_3|X_2) \leq I(X_1; Y_4|X_2), \quad I(X_2; Y_4|X_1) \leq I(X_2; Y_3|X_1)$$

for all $p(x_1)p(x_2)$. The capacity region of the interference channel in this regime is given by the set of rate pairs $(R_1, R_2)$ such that

$$R_1 \leq I(X_1; Y_3|X_2, Q),$$

$$R_2 \leq I(X_2; Y_4|X_1, Q),$$

$$R_1 + R_2 \leq \min\{I(X_1, X_2; Y_3|Q), \ I(X_1, X_2, Y_4|Q)\}$$

$$p(y_3, y_4|x_1, x_2)$$

Fig. 4. The 2-user interference channel.
for some $p(q)p(x_1|q)p(x_2|q)$.

The following proposition establishes the exponentially strong converse under strong interference. The strong converse for the interference channel with very strong interference (in addition to fixed-error second-order results) was derived in [26]. The strong converse for the Gaussian interference channel with strong interference was proved in [27].

**Proposition 14**: For an interference channel with strong interference, weak edge removal and the exponentially strong converse hold.

**Proof**: Note that the only nodes $i$ in an interference channel where $X_i \neq \emptyset$ are the encoder nodes, i.e. nodes 1 and 2. Thus, by Theorem 9 to prove the proposition it is enough to show that for any $k_n \in o(n)$, $R_{\{1,2\}}(N, 0^+, k_n) \subseteq R(N, 0^+)$, where $R(N, 0^+)$ is the region defined in (155)–(157).

We claim that an interference channel with strong interference also satisfies (154) for any joint distribution $p(x_1, x_2)$, even when $X_1, X_2$ are not independent. Indeed, for any $p(x_1, x_2)$, we have

$$I(X_1; Y_3|X_2) = \sum_{x_2} p(x_2) I(X_1; Y_3|X_2 = x_2)$$

$$\leq \sum_{x_2} p(x_2) I(X_1; Y_4|X_2 = x_2)$$

$$= I(X_1; Y_4|X_2)$$

where (159) follows from (154) since, conditioned on $X_2 = x_2$, $X_1$ is trivially independent from $X_2$. Thus, by the same proof as the lemma in [28] for the independent case, for any $p(x_1^n, x_2^n)$,

$$I(X_1^n; Y_3^n|X_2^n) \leq I(X_1^n; Y_4^n|X_2^n), \quad I(X_2^n; Y_4^n|X_1^n) \leq I(X_2^n; Y_3^n|X_1^n)$$

where

$$p(y_1^n, y_2^n|x_1^n, x_2^n) = \prod_{t=1}^n p_{Y_3,Y_4|X_1,X_2}(y_{1t}, y_{2t}|x_{1t}, x_{2t}).$$

Consider $(R_1, R_2) \in R_{\{1,2\}}(N, 0^+, k_n)$ where $k_n \in o(n)$. Let $Z_t$ be the signal sent on the edge $(a, b)$ at time $t$, so certainly $H(Z^n) \leq k_n$. Since for $j = 1, 2$, $X_j^n$ is a function of message $W_j$ and $Z^n$, we have

$$I(X_1^n; X_2^n|Z^n) \leq I(W_1; W_2|Z^n)$$

$$\leq I(W_1; W_2, Z^n)$$

$$= I(W_1; W_2) + I(W_1; Z^n|W_2)$$

(165)
\begin{align*}
&\leq H(Z^n) \\
&\leq k_n. 
\end{align*}

(166) 

(167)

Since node $a$ only has access to $W_1, W_2$, we have $(W_1, W_2, Z^n) → (X_1^n, X_2^n) → (Y_3^n, Y_4^n)$. We now write

\begin{align*}
nR_1 &= H(W_1 | W_2) \\
&= I(W_1; Y_3^n, Z^n | W_2) + H(W_1 | Y_3^n, W_2, Z^n) \tag{168} \\
&\leq I(W_1; Y_3^n | W_2, Z^n) + k_n + n\epsilon_n \tag{169} \\
&= I(W_1, X_1^n; Y_3^n | W_2, X_2^n, Z^n) + k_n + n\epsilon_n \tag{170} \\
&\leq I(X_1^n; Y_3^n | X_2^n, Z^n) + k_n + n\epsilon_n \tag{171} \\
&\leq I(X_1^n; Y_3^n) + k_n + n\epsilon_n, \tag{172}
\end{align*}

where in (170) we have used the fact that $H(Z^n) \leq k_n$, and Fano’s inequality, where $\epsilon_n \to 0$ as $n \to \infty$. Similarly

\begin{align*}
nR_2 &\leq nI(X_2^n, Y_4^n | X_1^n, Z^n) + k_n + n\epsilon_n. \tag{173}
\end{align*}

We also have

\begin{align*}
nR_1 &= H(W_1) \tag{174} \\
&\leq I(W_1; Y_3^n, Z^n) + n\epsilon_n \tag{175} \\
&\leq I(W_1; Y_3^n | Z^n) + k_n + n\epsilon_n \tag{176} \\
&\leq I(W_1, X_1^n; Y_3^n | Z^n) + k_n + n\epsilon_n \tag{177} \\
&= I(X_1^n; Y_3^n | Z^n) + I(W_1; Y_3^n | X_1^n, Z^n) + k_n + n\epsilon_n \tag{178} \\
&\leq I(X_1^n; Y_3^n | Z^n) + I(W_1; Y_3^n, X_2^n | X_1^n, Z^n) + k_n + n\epsilon_n \tag{179} \\
&= I(X_1^n; Y_3^n | Z^n) + I(W_1; X_2^n | X_1^n, Z^n) + k_n + n\epsilon_n \tag{180} \\
&\leq I(X_1^n; Y_3^n | Z^n) + I(W_1; W_2 | Z^n) + k_n + n\epsilon_n \tag{181} \\
&\leq I(X_1^n; Y_3^n | Z^n) + 2k_n + n\epsilon_n. \tag{182}
\end{align*}

Combining with one of the above steps for $R_2$ gives

\begin{align*}
n(R_1 + R_2) &\leq I(X_1^n; Y_3^n | Z^n) + I(X_2^n; Y_4^n | Z^n, X_1^n) + 3k_n + n2\epsilon_n \tag{183} \\
&\leq I(X_1^n; Y_3^n | Z^n) + I(X_2^n; Y_3^n | Z^n, X_1^n) + 3k_n + n2\epsilon_n \tag{184}
\end{align*}
where (184) follows from (161). To summarize,

\[ nR_1 \leq I(X_1^n; Y_3^n | X_2^n, Z^n) + k_n + n\epsilon_n, \]  
\[ nR_2 \leq I(X_2^n; Y_4^n | X_1^n, Z^n) + k_n + n\epsilon_n, \]  
\[ n(R_1 + R_2) \leq \min\{I(X_1^n, X_2^n; Y_3^n | Z^n), I(X_1^n, X_2^n; Y_4^n | Z^n)\} + 3k_n + n2\epsilon_n, \]  
\[ k_n \geq I(X_1^n; X_2^n | Z^n). \]

One can see that this is precisely the region for the interference channel when both messages are required to be decoded at both decoders, except that we have close-to-independence instead of exact independence. The difficulty with condition (189) is not just that \( X_1^n, X_2^n \) are not perfectly independent, but that the dependence between individual letters \( X_{1t}, X_{2t} \) may varying depending on \( t \). The method of Dueck in [29] (also similar to Ahlswede’s “wringing” technique [30]) allows us to show that for most \( t \in [1 : n] \), the letters \( X_{1t}, X_{2t} \) are nearly independent. This will allow single-letterization of the region in (186)–(189). In particular, there exist some \( m \leq \sqrt{nk_n} \) and \( t_1, \ldots, t_m \in [1 : n] \), where for all \( t \in [1 : n] \)

\[ I(X_{1t}; X_{2t}|Q') \leq \sqrt{\frac{k_n}{n}}. \]

where

\( Q' = (Z^n, X_{1t_1}, \ldots, X_{1t_m}, X_{2t_1}, \ldots, X_{2t_m}). \)

We reproduce the essential proof of this fact from [29] as follows. First, if there is no \( t \) where \( I(X_{1t}; X_{2t}|Z^n) > \sqrt{\frac{k_n}{n}} \), then we are done. Otherwise, take \( t_1 \) to be one such \( t \). We may write

\[ I(X_1^n; X_2^n | Z^n, X_{1t_1}, X_{2t_1}) = I(X_1^n; X_2^n | Z^n) - I(X_1^n; X_{2t_1} | Z^n) - I(X_{1t_1}; X_{2t_1} | Z^n), \]

\[ \leq I(X_1^n; X_2^n | Z^n) - I(X_{1t_1}; X_{2t_1} | Z^n) \]

\[ < k_n - \sqrt{\frac{k_n}{n}}. \]

Now, if there is no \( t \) where \( I(X_{1t}; X_{2t}|Z^n, X_{1t_1}, X_{2t_1}) > \sqrt{\frac{k_n}{n}} \), then again we are done. Otherwise, take \( t_2 \) to be one such \( t \), and proceed as above. This process must terminate after a finite number (say \( m \)) of steps, at which point (190) must hold for all \( t \), and by a similar argument as in (192)–(194), we have

\[ I(X_1^n; X_2^n | Q') \leq k_n - m\sqrt{\frac{k_n}{n}}. \]
Since the mutual information is nonnegative, we have \( m \leq \sqrt{n k_n} \).

We now have

\[
I(X^n_1; Y^n_3 | X^n_2, Z^n) \leq I(X^n_1; Y^n_3 | X^n_2, Q') + H(X_{1t_1}, \ldots, X_{1t_m}, X_{2t_1}, \ldots, X_{2t_m})
\]

\[
\leq I(X^n_1; Y^n_3 | X^n_2, Q') + m \log |\mathcal{X}_1| \cdot |\mathcal{X}_2|
\]

\[
\leq I(X^n_1; Y^n_3 | X^n_2, Q') + \sqrt{n k_n} \log |\mathcal{X}_1| \cdot |\mathcal{X}_2|
\]

\[
= \sum_{t=1}^{n} I(X^n_1; Y^n_3 | X^n_2, Y_{3t-1}, X^n_2, Q') + n \sqrt{\delta} \log |\mathcal{X}_1| \cdot |\mathcal{X}_2|
\]

\[
\leq \sum_{t=1}^{n} I(X^n_1; Y^n_3 | X^n_2, Q') + n \sqrt{\delta} \log |\mathcal{X}_1| \cdot |\mathcal{X}_2|
\]

\[
= n I(X_1; Y_3 | X_2, Q) + n \sqrt{\delta} \log |\mathcal{X}_1| \cdot |\mathcal{X}_2|
\]

where

\( Q'' \sim \text{Unif}[1 : n], \quad Q = (Q', Q''), \quad X_1 = X_{1Q''}, \quad X_2 = X_{2Q''}, \quad Y_3 = Y_{3Q''}, \quad Y_4 = Y_{4Q''} \).  

(202)

Applying (186), and performing similar analyses for (187)–(188), combined with (190), we have

\[
R_1 \leq I(X_1; Y_3 | X_2, Q) + \frac{k_n}{n} + \epsilon_n + \sqrt{\frac{k_n}{n}} \log |\mathcal{X}_1| \cdot |\mathcal{X}_2|,
\]

(203)

\[
R_2 \leq I(X_2; Y_4 | X_1, Q) + \frac{k_n}{n} + \epsilon_n + \sqrt{\frac{k_n}{n}} \log |\mathcal{X}_1| \cdot |\mathcal{X}_2|,
\]

(204)

\[
R_1 + R_2 \leq \min\{I(X_1, X_2; Y_3 | Q), \quad I(X_1, X_2, Y_4 | Q)\} + \frac{3k_n}{n} + 2 \epsilon_n + \sqrt{\frac{k_n}{n}} \log |\mathcal{X}_1| \cdot |\mathcal{X}_2|,
\]

(205)

\[
\sqrt{\frac{k_n}{n}} \geq I(X_1; X_2 | Q).
\]

(206)

Recalling that \( k_n \in o(n) \) and \( \epsilon_n \to 0 \), and using the fact that mutual information is continuous for fixed finite alphabets, taking a limit as \( n \to \infty \) yields the region from (155)–(157).

VIII. Conclusions

This paper explored the relationship between edge removal properties and strong converses. Our main results are summarized in Fig. 1. We found three main levels of properties for both edge removal and strong converse, and showed that for a very large class of networks, the strong converse property implies the corresponding edge removal property. We found that implications in
the opposite direction hold for deterministic networks and sometimes for memoryless stationary networks.

Our strongest results are those for the “middle” level, connecting the weak edge removal property to the exponentially strong converse. In particular, we showed that these properties are equivalent for all discrete memoryless stationary networks. Thus, if an existing weak converse or outer bound can be strengthened to show that it still holds in the presence of an extra link carrying a sub-linear number of bits, then the converse or outer bound also holds in an exponentially strong sense, meaning that for any rate vector outside the region, the probability of error converges to 1 exponentially fast. It appears that many existing arguments can be strengthened in this sense with relatively little effort, thereby proving exponentially strong results. We believe that this middle level deserves more focus than it has received, because exponentially strong converses and weak edge removal properties seem to hold for so many problems (at least under average probability of error). Therefore, one should always ask whether a given converse proof can be strengthened in this sense.

Several open problems remain. The most important is whether edge removal and strong converse properties hold in general. In particular, we know of no memoryless stationary network for which the weak edge removal property or the exponentially strong converse does not hold under average probability of error. The techniques of Sec. VII seem to allow one to prove a weak edge removal property (and thus an exponentially strong converse) for most (perhaps all) existing single-letter outer bounds, but there is no apparent way to do this without an existing single-letter result. Our observation that the properties hold for the discrete broadcast channel suggest that it may be possible to prove such results even for problems with unknown capacity regions, but we know of no other cases where this has been done.

Secondly, we conjecture that an equivalence holds for discrete memoryless networks on the “lower layer” in Fig. I between very weak edge removal and the ordinary strong converse, but we have only been able to prove this result for deterministic networks and acyclic networks of independent point-to-point links. Thirdly, most of our results showing that edge removal properties imply strong converses hold only for discrete systems; generalizing this direction to continuous systems and general alphabets would significantly improve the results. Finally, it would be interesting to find a strong converse property equivalent to the extremely weak edge removal property.
APPENDIX A
PROOF OF PROPOSITION [1]

We will show that $\mathcal{R}(\mathcal{N}, \epsilon_n) \subseteq \mathcal{R}(\mathcal{N}, \tilde{\epsilon}_n)$; the opposite direction follows by reversing the roles of $\epsilon_n$ and $\tilde{\epsilon}_n$. Fix any rate vector

$$R \in \bigcup_{n_0 \in \mathbb{N}} \bigcap_{n \geq n_0} \mathcal{R}^{(n)}(\mathcal{N}, \epsilon_n).$$

(207)

We aim to show that $R \in \mathcal{R}(\mathcal{N}, \tilde{\epsilon}_n)$. There exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $R \in \mathcal{R}^{(n)}(\mathcal{N}, \epsilon_n)$. By the assumption of the lemma, there exists a subsequence $n_i$ such that

$$\lim_{i \to \infty} -\frac{1}{n_i} \log(1 - \epsilon_{n_i}) = \delta.$$  (208)

For sufficiently large $i$, we have $n_i \geq n_0$, so $R \in \mathcal{R}^{(n_i)}(\mathcal{N}, \epsilon_{n_i})$. That is, there exists an $n_i$-length code with rate $R$ and probability of error at most $\epsilon_{n_i}$. Fix integer $N$, and form a new code on network $\mathcal{N}$ of length $n_iN$ and rate $\frac{N-2}{N}R$ as follows. Roughly, we repeat the original code $N$ times, and use an outer MDS code for a small amount of error correction. In particular, for each node $v \in [1 : d]$ where $R_v > 0$, form a $(N, N-2)$ MDS code on symbols from the finite field of order $2^\lfloor n_i R_v \rfloor$. This code exists for sufficiently large $i$. Let the MDS codeword be denoted by $(W_v(1), \ldots, W_v(N))$. Repeat the original code $N$ times, where on the $\ell$th repetition $W_v(\ell)$ is treated as the message originating at node $v$. Because each outer code is MDS, one error can be corrected, so if it most one of the $N$ repetitions results in an error, the full code will decode correctly. Because the network is memoryless and stationary, each repetition is independent and results in error with probability $\epsilon_{n_i}$, so the probability of error for the full code is given by

$$P_e = 1 - (1 - \epsilon_{n_i})^N + N \epsilon_{n_i} (1 - \epsilon_{n_i})^{N-1}$$  (209)

$$= 1 - (1 - \epsilon_{n_i})^{N-1} [1 - \epsilon_{n_i} + N \epsilon_{n_i}].$$  (210)

Note that $208^\delta$ and the assumption that $\delta > 0$ imply that $\epsilon_{n_i} \to 1$, meaning $1 - \epsilon_{n_i} + N \epsilon_{n_i} \to N$. Thus

$$\lim_{i \to \infty} \frac{1}{n_i} \log(1 - P_e) = \lim_{i \to \infty} \frac{1}{n_i} [\log(1 - \epsilon_{n_i}) + N]$$

$$= -(N - 1)\delta.$$  (211)

(212)

In particular, for sufficiently large $i$, we have

$$1 - P_e \geq \exp\{-n_i(N - 1/2)\delta\}$$  (213)
Hence, for any \( N \) and sufficiently large \( i \),
\[
\frac{N - 2}{N^2} R \in \mathcal{R}(n_i N)(N, 1 - \exp\{-n_i(N - 1/2)\delta\}). \tag{214}
\]

Consider any blocklength \( m \) where \( n_i N \leq m \leq n_i(N + 1) \). We may convert a code with blocklength \( n_i N \) to one with blocklength \( m \) simply by ignoring the additional \( m - n_i N \) symbols. This reduces the rate by a factor of \( \frac{n_i N}{m} \geq \frac{N}{N+1} \), but does not change the probability of error. Thus we have
\[
\frac{N - 2}{N + 1} R \in \mathcal{R}(m)(N, 1 - \exp\{-n_i(N - 1/2)\delta\}). \tag{215}
\]

By the liminf assumption on \( \bar{\epsilon}_n \), for sufficiently large \( m \) we have
\[
-\frac{1}{m} \log(1 - \bar{\epsilon}_m) \geq \frac{N - 1/2}{N} \delta. \tag{216}
\]

Thus, if \( m \geq n_i N \), we have
\[
\bar{\epsilon}_m \geq 1 - \exp\left\{-m \frac{N - 1/2}{N} \delta\right\}, \tag{217}
\]
\[
\geq 1 - \exp\{-n_i(N - 1/2)\delta\} \tag{218}
\]
where (217) holds by (216) for sufficiently large \( i \). Hence, for any \( N \), for all \( m \) sufficiently large we have
\[
\frac{N - 2}{N + 1} R \in \mathcal{R}(m)(N, \bar{\epsilon}_m). \tag{219}
\]
Thus
\[
\frac{N - 2}{N + 1} R \in \mathcal{R}(N, \bar{\epsilon}_n). \tag{220}
\]

As this holds for all \( N \), by closure we have \( R \in \mathcal{R}(N, \bar{\epsilon}_n) \).

**APPENDIX**

**PROOF OF PROPOSITION 2**

To prove the claim, first note that
\[
\max_{q(x), v(y|x)} I_{q \times v}(X; Y) = \min\{|\mathcal{X}|, |\mathcal{Y}|\}. \tag{221}
\]

Let \( \bar{q}(x), \bar{v}(y|x) \) achieve the maximum in (221). Also let \( q^*(x) \) be a capacity-achieving input distribution for the channel \( p(y|x) \). If \( C < \log \min\{|\mathcal{X}|, |\mathcal{Y}|\} \), then \( p(y|x) \neq \bar{v}(y|x) \). Additionally, for any \( C < R < \log \min\{|\mathcal{X}|, |\mathcal{Y}|\} \), by continuity of mutual information there exists \( \lambda \in (0, 1) \) such that
\[
R = I_{q_\lambda \times v_\lambda}(X; Y) \tag{222}
\]
where

\[ q_\lambda(x) = (1 - \lambda)q^*(x) + \lambda \bar{q}(x), \tag{223} \]
\[ v_\lambda(y|x) = (1 - \lambda)p(y|x) + \lambda \bar{v}(y|x). \tag{224} \]

Thus

\[ \alpha(R) \leq D(v_\lambda(y|x) || p(y|x)|q_\lambda(x)). \tag{225} \]

Noting that as \( R \) approaches \( C \), \( \lambda \) approaches 0, so \( v_\lambda(y|x) \) approaches \( p(y|x) \). As the relative entropy behaves quadratically when its arguments are close together \[\text{[31]}\], we see that for \( R \) close to \( C \), \( \alpha'(C) = 0 \). Thus \( \alpha'(C) = 0 \).

Now consider a channel where \( C = \log \min(|\mathcal{X}|, |\mathcal{Y}|) \). Thus, if \( R > C \), it must be that \( R > I_{q \times v}(X;Y) \) for any choice of \( q(x) \) and \( v(y|x) \), so the second term in \[\text{[12]}\] is positive. In particular,

\[ \alpha(R) = \min_{q(x)} \min_{v(y|x)} D(v(y|x) || p(y|x)|q(x)) + R - I_{q \times v}(X;Y) \tag{226} \]
\[ \geq \min_{q(x)} \min_{v(y|x)} R - I_{q \times v}(X;Y) \tag{227} \]
\[ = R - C. \tag{228} \]

Moreover, this lower bound is achievable simply by setting \( q(x) = q^*(x) \), and \( v(y|x) = p(y|x) \).

**Appendix C**

**Proof of Proposition** [3]

To prove statement 1 of the proposition, observe that

\[ \bigcup_{k \in \mathbb{N}} \mathcal{R}(\mathcal{N}, 0^+, k) \subseteq \bigcap_{k_n \to \infty} \mathcal{R}(\mathcal{N}, 0^+, k_n) \tag{229} \]
\[ \subseteq \bigcup_{k_n \in o(n)} \mathcal{R}(\mathcal{N}, 0^+, k_n) \tag{230} \]
\[ \subseteq \bigcap_{\delta > 0} \mathcal{R}(\mathcal{N}, 0^+, \delta n) \tag{231} \]

where \[\text{[229]}\] holds since for any \( k \in \mathbb{N} \) and sequence \( k_n \to \infty \), \( k \leq k_n \) for sufficiently large \( n \); \[\text{[230]}\] holds because the intersection in \[\text{[229]}\] includes at least one sequence in \( o(n) \); and \[\text{[231]}\] holds because \( k_n \leq \delta n \) for sufficiently large \( n \) for any \( \delta > 0 \) and \( k_n \in o(n) \). Note that if strong edge removal holds, for any \( R \) in the RHS of \[\text{[231]}\], \( R - K\delta \in \mathcal{R}(\mathcal{N}, 0^+) \) for all \( \delta > 0 \).
Since $\mathcal{R}(\mathcal{N}, 0^+]$ is closed, this implies $\mathbb{R} \in \mathcal{R}(\mathcal{N}, 0^+)$. Therefore, since the RHS of (230) is contained in the RHS of (231), strong edge removal implies weak edge removal. Since the RHS of (229) is contained in the RHS of (230), weak edge removal implies very weak edge removal. Similarly, since the LHS of (229) is contained in the RHS of (229), very weak edge removal implies extremely weak edge removal.

To prove statement 2 it is enough to show that for all $\epsilon > 0$,

$$
\bigcap_{k_n: k_n \to \infty} \mathcal{R}(\mathcal{N}, \epsilon, k_n) = \bigcup_{k \in \mathbb{N}} \mathcal{R}(\mathcal{N}, \epsilon, k).
$$

(232)

For any $k \in \mathbb{N}$ and any sequence $k_n \to \infty$, $k \geq k_n$ for sufficiently large $n$. Thus

$$
\bigcap_{k_n: k_n \to \infty} \mathcal{R}(\mathcal{N}, \epsilon, k_n) \supseteq \bigcup_{k \in \mathbb{N}} \mathcal{R}(\mathcal{N}, \epsilon, k).
$$

(233)

Taking a closure yields $\supseteq$ in (232), since the LHS of (232) is already closed. To prove the opposite direction, let $\gamma_k$ be a positive sequence where $\lim_{k \to \infty} \gamma_k \to 0$. Recall that for any $k \in \mathbb{N}$,

$$
\mathcal{R}(\mathcal{N}, \epsilon, k) = \bigcup_{n_0 \in \mathbb{N}} \bigcap_{n \geq n_0} \mathcal{R}(n, \epsilon, k).
$$

(234)

In particular, for any $k$ there exists $n_0(k)$ such that for all $n \geq n_0(k)$, we have

$$
\mathcal{R}(n, \epsilon, k) \subseteq \mathcal{R}(\mathcal{N}, \epsilon, k) + \gamma_k.
$$

(235)

Now define a sequence

$$
k_n = \max\{k : n \geq n_0(k)\}.
$$

(236)

Note that $k_n \to \infty$ as $n \to \infty$, because for any $k$, $k_n \geq k$ for all $n \geq n_0(k)$. Thus the LHS of (232) is contained in $\mathcal{R}(\mathcal{N}, \epsilon, k_n)$. Moreover

$$
\mathcal{R}(\mathcal{N}, \epsilon, k_n) = \bigcup_{n_0 \in \mathbb{N}} \bigcap_{n \geq n_0} \mathcal{R}(n, \epsilon, k_n)
$$

(237)

$$
\subseteq \bigcup_{n_0 \in \mathbb{N}} \bigcap_{n \geq n_0} (\mathcal{R}(\mathcal{N}, \epsilon, k_n) + \gamma_{k_n})
$$

(238)

$$
= \bigcup_{n_0 \in \mathbb{N}} \bigcap_{n \geq n_0} \mathcal{R}(\mathcal{N}, \epsilon, k_n)
$$

(239)

$$
\subseteq \bigcup_{n_0 \in \mathbb{N}} \mathcal{R}(\mathcal{N}, \epsilon, k_{n_0})
$$

(240)

$$
\subseteq \bigcup_{k \in \mathbb{N}} \mathcal{R}(\mathcal{N}, \epsilon, k)
$$

(241)
where (237) holds by definition, (238) follows from (235), and (239) holds because $\gamma_k \to 0$. This proves $\subseteq$ in (232).

To prove statement 3, we will show that

$$\bigcup_{k_n \in o(n)} \mathcal{R}(N', 0^+, k_n) = \bigcap_{\delta > 0} \mathcal{R}(N', 0^+, \delta n).$$  \hspace{1cm} (242)

The $\subseteq$ direction holds by the argument in (230)–(231). To prove the opposite direction, let $R$ be in the RHS of (242). Thus, for all $\epsilon, \delta, \gamma > 0$, for sufficiently large $n$ we have $R - \gamma \in \mathcal{R}^{(n)}(N', \epsilon, n\delta)$. In particular, if we let $\gamma_r$ be a positive sequence converging to 0, then for all $r$ there exists $n_0(r)$ such that for all $n \geq n_0(r)$ we have

$$R - \gamma_r \in \mathcal{R}^{(n)}(N, 1/r, n/r).$$  \hspace{1cm} (243)

Let

$$k_n = \min \{n/r : n \geq n_0(r)\}.$$  \hspace{1cm} (244)

Note that for all $n \geq n_r$ we have $k_n \leq n/r$; thus $k_n \in o(n)$. Moreover, for all $n$ we have $n \geq n_0(n/k_n)$, so

$$R - \gamma_{n/k_n} \in \mathcal{R}^{(n)}(N, k_n/n, k_n).$$  \hspace{1cm} (245)

Since $k_n/n \to 0$, we have $\gamma_{n/k_n} \to 0$, so $R \in \mathcal{R}(N', 0^+, k_n)$. This proves $\supseteq$ in (242).

**APPENDIX D**

**PROOF OF THEOREM 13**

A significant technical tool in proving network equivalence is the idea of channel simulation, in which a point-to-point channel is accurately simulated by any other with higher capacity. This idea was at the heart of the proof in [25]. A version of this idea was stated in [32] as the universal channel simulation lemma, stated as follows. While [32] did not provide a proof, we presented a proof in the appendix of [33].

**Lemma 15:** Let $(\mathcal{X}, q(y|x), \mathcal{Y})$ be a discrete memoryless channel with capacity $C$. Given a rate $R > C$, a channel simulation code $(f, g)$ consists of

- $f : \mathcal{X}^n \times [0, 1] \to \{0, 1\}^{nR}$,
- $g : \{0, 1\}^{nR} \times [0, 1] \to \mathcal{Y}^n$.

Let $p(y^n|x^n)$ be the conditional pmf of $Y^n$ given $X^n$ where $U \sim \text{Unif}[0, 1]$ and

$$Y^n = g(f(X^n, U), U).$$  \hspace{1cm} (246)
There exists a sequence of length-$n$ simulation codes where

$$\lim_{n \to \infty} \max_{x^n} d_{TV}(p(y^n|x^n), q(y^n|x^n)) = 0. \quad (247)$$

We now proceed to prove Theorem 13. By Theorem 4, we only need to show that the very weak edge removal property implies the ordinary strong converse. The basic approach is to use network equivalence to convert a code for noisy network $N$ into a code on the noiseless version, then apply Lemma 8 on this noiseless network, and then again use network equivalence to convert back to the noisy network.

Let $E \subset [1 : d] \times [1 : d]$ be the set of pairs of nodes connected by point-to-point links. Recall that by assumption, the directed graph $([1 : d], E)$ is acyclic. Thus we may assign each node $i$ a distinct integer $\pi_i \in [1 : d]$ where $\pi_i < \pi_j$ if $(i, j) \in E$. For any $(i, j) \in E$, let $C_{i\to j}$ be the capacity of the link from $i$ to $j$. Assume without loss of generality that $C_{i\to j} > 0$ for all $(i, j) \in E$. Let $C_{\min} = \min_{(i,j) \in E} C_{i\to j}$, so in particular $C_{\min} > 0$. Denote $X_{i\to j}$ and $Y_{i\to j}$ as the input and output respectively of the link $(i, j)$. Thus the transmitted symbol from node $i$ can be written

$$X_i = (X_{i\to j} : (i, j) \in E) \quad (248)$$

and the received symbol at node $j$ can be written

$$Y_j = (Y_{i\to j} : (i, j) \in E). \quad (249)$$

Let $R$ be achievable with respect to fixed $\epsilon \in (0, 1)$. Thus, for sufficiently large $n$, there exists a length-$n$ code for network $N$ with rate $R$ and probability of error $\epsilon$. Let $\phi_{it}$ be the encoding function for node $i$ and time $t$, and $\psi_i$ the decoding function at node $i$. Define the block-wise encoding function at node $i$

$$\phi^n_i : [1 : 2^{nR_i}] \times Y^n_i \to X^n_i \quad (250)$$
as

$$\phi^n_i(w, y^n_i) = (\phi_{i1}(w), \phi_{i2}(w, y_{i1}), \ldots, \phi_{im}(w, y^n_{i1})). \quad (251)$$

Using the notation in (249), we may write

$$\phi^n_i(w, y^n_{i\to j} : (i, j) \in E). \quad (252)$$

Due to the network being acyclic, we may form a pipelined block-Markov version of this code as follows. Given integer $N$, we form a code with length $n(N + d)$ and rate $\frac{N}{N+d}R$. The outer
blocklength $N$ serves a similar function as it did for network stacking, but here it represents the number of message blocks transmitted subsequently, rather than the number of stacks. Note that message $i$ consists of $nR_i$ bits, which we denote $W_i(1), \ldots, W_i(N)$, each consisting of $nR_i$ bits. We then pipeline $N$ copies of the original code, encoding $n$-length blocks at a time. In particular, we introduce notation

\[ X_j^{n(N+d)}(\ell + \pi_j) = \phi^n_j(W_j(\ell), Y_{i\rightarrow j}^n(\ell + \pi_i) : (i, j) \in \mathcal{E}). \]  

(255)

Now, we define the coding operations at node $j$ by, for all $\ell \in [1 : N]$,

\[ X_j^n(\ell + \pi_j) = \phi^n_j(W_j(\ell), Y_{i\rightarrow j}^n(\ell + \pi_i) : (i, j) \in \mathcal{E}). \]  

(255)

Recall that if $(i, j) \in \mathcal{E}$, then $\pi_i < \pi_j$, meaning that the arguments of $\phi^n_j$ in (255) are causally available. Moreover, $\ell + \pi_i \leq N + d$, so the blocklength of $n(N + d)$ is respected. Note that (255) does not specify all channel inputs, namely $X_j^n(\ell')$ for $\ell' < 1 + \pi_j$ or $\ell' > N + \pi_j$; these channel inputs can be arbitrary, as the corresponding channel outputs will be ignored. To decode at node $i$, for all $\ell \in [1 : N]$ let

\[ (\hat{W}_{ji}(\ell) : i \in D_j) = \psi_i(W_i(\ell), Y_{k\rightarrow i}^n(\ell + \pi_k) : (k, i) \in \mathcal{E}). \]  

(256)

Observe that the variables associated with a given index $\ell \in [1 : N]$ associate only with themselves, and behave exactly like the original $n$-length code. Thus, an error occurs on this pipelined code if and only if any of the $N$ copies make an error, so the probability of error is

\[ 1 - (1 - \epsilon)^N. \]  

(257)

Thus we have

\[ \frac{N}{N + d} \mathbf{R} \in \mathcal{R}^{(n(N+d))}(\mathcal{N}, 1 - (1 - \epsilon)^N). \]  

(258)

Note that in this pipelined code, encoding operations are performed on $n$-length blocks at a time. Thus, the pipelined code on $\mathcal{N}$ can be converted to one on a deterministic network using channel simulation codes. In particular, fix $\Delta \in (0, C_{\text{min}})$ and let $\mathcal{N}_\Delta$ be the network of noiseless links where link $(i, j)$ is replaced by a noiseless link with capacity $C_{i\rightarrow j} + \Delta$. By Lemma 15, for each link $(i, j)$ there exists a channel simulation code for link $(i, j)$ of rate $C_{i\rightarrow j} + \Delta$ and total variational distance at most $d_n^{(i\rightarrow j)}$, where $d_n^{(i\rightarrow j)} \rightarrow 0$ as $n \rightarrow \infty$. We use $N$ copies of this channel simulation code to simulate the behavior of link $(i, j)$ in network $\mathcal{N}$ using the corresponding link.
on \( \tilde{N}_{\Delta} \). Note that the bound (247) on the variational distance of the simulated code applies for any input sequence to the channel, so \( N \) copies of the channel simulation code for \((i, j)\) increases the probability of error by at most \( Nd_n^{(i\rightarrow j)} \). Thus, the probability of error of the resulting code on \( \tilde{N}_{\Delta} \) is at most

\[
1 - (1 - \epsilon)^N + \sum_{(i,j) \in E} Nd_n^{(i\rightarrow j)} \leq 1 - \frac{1}{2}(1 - \epsilon)^N
\]  

(259)

where the inequality holds for sufficiently large \( n \), since each sequence \( d_n^{(i\rightarrow j)} \) vanishes with \( n \). Recall that the channel simulation codes described in Lemma 15 employ common randomness \( U \) between the transmitter and receiver of each link. Thus, a direct application of Lemma 15 implies only the existence of a code achieving the probability in (259) if nodes are allowed common randomness. However, we may treat this common randomness as a randomized codebook, and employ a usual random coding argument to show that there exists at least one deterministic code achieving (259). Hence, for sufficiently large \( n \),

\[
\frac{N}{N + d} \mathcal{R} \in \mathcal{R}^{(n(N + d))}(\tilde{N}_{\Delta}, 1 - \frac{1}{2}(1 - \epsilon)^N).
\]

(260)

We now apply Lemma 8 on \( \tilde{N}_{\Delta} \), to find that there exists a constant \( K \) such that for any \( \tilde{\epsilon} > 0 \), for sufficiently large \( n \) we have

\[
\frac{N}{N + d} \mathcal{R} \in \mathcal{R}^{(n(N + d))}(\tilde{N}_{\Delta}, \tilde{\epsilon}, \eta(\tilde{\epsilon}) - KN \log(1 - \epsilon) + K).
\]

(261)

Let \( \tilde{N}_{-\Delta} \) be the noiseless network where each link \((i, j)\) is replaced by a noiseless one with capacity \( C_{i\rightarrow j} - \Delta \). By the assumption that \( \Delta < C_{\text{min}} \), we always have \( C_{i\rightarrow j} - \Delta > 0 \). We may convert the code on \( \tilde{N}_{\Delta} \) to one on \( \tilde{N}_{-\Delta} \) by stretching each block of \( n \) to one of length

\[
n' = \frac{C_{\text{min}} + \Delta}{C_{\text{min}} - \Delta} n.
\]

(262)

Thus

\[
\frac{N}{N + d} \cdot \frac{C_{\text{min}} - \Delta}{C_{\text{min}} + \Delta} \mathcal{R} \in \mathcal{R}^{(n'(N + d))}(\tilde{N}_{-\Delta}, \tilde{\epsilon}, \eta(\tilde{\epsilon}) - K N \log(1 - \epsilon) + K).
\]

(263)

Now we use ordinary noisy channel codes to convert this code back to one on \( \tilde{N} \), again one block (now of length \( n' \)) at a time. For any \( N \) and sufficiently large \( n \), the probability of an error occurring on any of these channel codes can be made at most \( \tilde{\epsilon} \). Thus we have

\[
\frac{N}{N + d} \cdot \frac{C_{\text{min}} - \Delta}{C_{\text{min}} + \Delta} \mathcal{R} \in \mathcal{R}^{(n'(N + d))}(N, 2\tilde{\epsilon}, \eta(2\tilde{\epsilon}) - K N \log(1 - \epsilon) + K).
\]

(264)

As the above holds for any \( \tilde{\epsilon} > 0 \), we may write

\[
\frac{N}{N + d} \cdot \frac{C_{\text{min}} - \Delta}{C_{\text{min}} + \Delta} \mathcal{R} \in \bigcap_{\tilde{\epsilon} > 0} \mathcal{R}(N, 2\tilde{\epsilon}, \eta(2\tilde{\epsilon}) - K N \log(1 - \epsilon) + K).
\]

(265)
\[ \subseteq \bigcap \bigcup_{\varepsilon > 0} k \in \mathbb{N} \mathcal{R}(\mathcal{N}, \varepsilon, k). \tag{266} \]

Since we may take \( N \) to be arbitrarily large, and \( \Delta \) arbitrarily small, and we chose \( R \) to be any achievable vector with respect to \( \varepsilon \), by closure we have

\[ \mathcal{R}(\mathcal{N}, \varepsilon) \subseteq \bigcap \bigcup_{\varepsilon > 0} k \in \mathbb{N} \mathcal{R}(\mathcal{N}, \varepsilon, k). \tag{267} \]

By Proposition 3, very weak edge removal property if and only if the right-hand side of (267) equals \( \mathcal{R}(\mathcal{N}, 0^+) \). Therefore, the very weak edge removal property implies the strong converse.

**REFERENCES**

[1] T. Ho, M. Effros, and S. Jalali, “On equivalence between network topologies,” in *Proc. Forty-Eighth Annual Allerton Conference*, Monticello, IL, Oct. 2010.

[2] S. Jalali, M. Effros, and T. Ho, “On the impact of a single edge on the network coding capacity,” in *Proc. Information Theory and Applications Workshop (ITA)*, San Diego, CA, Feb. 2011, pp. 1–5.

[3] E. J. Lee, M. Langberg, and M. Effros, “Outer bounds and a functional study of the edge removal problem,” in *Proc. IEEE Information Theory Workshop*, Sevilla, Spain, Sep. 2013, pp. 1–5.

[4] S. U. Kamath, D. N. C. Tse, and V. Anantharam, “Generalized network sharing outer bound and the two-unicast problem,” in *Proc. International Symposium on Network Coding (NetCod)*, Beijing, China, Jul. 2011.

[5] R. W. Yeung, “A framework for linear information inequalities,” *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1924–1934, Nov. 1997.

[6] M. Langberg and M. Effros, “Network coding: Is zero error always possible?” in *Proc. Forty-Nine Annual Allerton Conference*, Monticello, IL, Sep. 2011, pp. 1–8.

[7] T. H. Chan and A. Grant, “Network coding capacity regions via entropy functions,” *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5347–5374, Sep. 2014.

[8] M. F. Wong, M. Langberg, and M. Effros, “On a capacity equivalence between network and index coding and the edge removal problem,” in *Proc. IEEE International Symposium on Information Theory*, Istanbul, Turkey, Jun. 2013, pp. 972–976.

[9] P. Noorzad, M. Effros, M. Langberg, and T. Ho, “On the power of cooperation: Can a little help a lot?” in *2014 IEEE International Symposium on Information Theory*, June 2014, pp. 3132–3136.

[10] P. Noorzad, M. Effros, and M. Langberg, “On the cost and benefit of cooperation,” in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 36–40.

[11] ———, “Can negligible cooperation increase network reliability?” in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1784–1788.

[12] ———, “The unbounded benefit of encoder cooperation for the \( k \)-user MAC,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 340–344.

[13] ———, “Can negligible rate increase network reliability?” *IEEE Transactions on Information Theory*, vol. PP, no. 99, pp. 1–1, 2017.

[14] ———, “The benefit of encoder cooperation in the presence of state information,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017.
[15] M. Langberg and M. Effros, “On the capacity advantage of a single bit,” in *2016 IEEE Globecom Workshops (GC Wkshps)*, Dec 2016, pp. 1–6.

[16] W. Gu, “On achievable rate regions for source coding over networks,” Ph.D. dissertation, California Institute of Technology, 2009.

[17] S. L. Fong and V. Y. F. Tan, “Strong converse theorems for multimeasure networks with tight cut-set bound,” [Online] arXiv:1606.04678, Jun. 2016.

[18] K. Marton, “A simple proof of the blowing-up lemma (corresp.),” *IEEE Transactions on Information Theory*, vol. 32, no. 3, pp. 445–446, May 1986.

[19] S. Arimoto, “On the converse to the coding theorem for discrete memoryless channels (corresp.),” *IEEE Transactions on Information Theory*, vol. 19, no. 3, pp. 357–359, May 1973.

[20] G. Dueck and J. Körner, “Reliability function of a discrete memoryless channel at rates above capacity (corresp.),” *Information Theory, IEEE Transactions on*, vol. 25, no. 1, pp. 82–85, Jan 1979.

[21] M. Langberg and M. Effros, “Source coding for dependent sources,” in *Information Theory Workshop (ITW), 2012 IEEE*, Sept 2012, pp. 70–74.

[22] R. Ahlswede, P. Gács, and J. Körner, “Bounds on conditional probabilities with applications in multi-user communication,” *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, pp. 157–177, 1976.

[23] M. Raginsky and I. Sason, “Concentration of measure inequalities in information theory, communications, and coding,” *Foundations and Trends in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.

[24] V. Strassen, “The existence of probability measures with given marginals,” *Ann. Math. Statist.*, vol. 36, pp. 423–439, 1965.

[25] R. Koetter, M. Effros, and M. Medard, “A theory of network equivalence— part i: Point-to-point channels,” *IEEE Trans. on Information Theory*, vol. 57, no. 2, pp. 972–995, 2011.

[26] S. Q. Le, V. Y. F. Tan, and M. Motani, “A case where interference does not affect the channel dispersion,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2439–2453, May 2015.

[27] S. L. Fong and V. Y. F. Tan, “A proof of the strong converse theorem for gaussian multiple access channels,” *IEEE Transactions on Information Theory*, vol. 62, no. 8, pp. 4376–4394, Aug 2016.

[28] M. Costa and A. E. Gamal, “The capacity region of the discrete memoryless interference channel with strong interference (corresp.),” *IEEE Transactions on Information Theory*, vol. 33, no. 5, pp. 710–711, Sep 1987.

[29] G. Dueck, “The strong converse to the coding theorem for the multiple-access channel,” *J. Combinat., Inf. Syst. Sci.*, vol. 6, no. 3, pp. 187–196, 1981.

[30] R. Ahlswede, “An elementary proof of the strong converse theorem for the multiple access channel,” *J. Combinat., Inf. Syst. Sci.*, vol. 7, no. 3, pp. 216–230, 1982.

[31] S. Borade and L. Zheng, “Euclidean information theory,” in *2008 IEEE International Zurich Seminar on Communications*, March 2008, pp. 14–17.

[32] Y. Xiang and Y.-H. Kim, “A few meta-theorems in network information theory,” in *Information Theory Workshop (ITW), 2014 IEEE*, Nov 2014, pp. 77–81.

[33] O. Kosut and J. Kliewer, “Equivalence for networks with adversarial state,” *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4137–4154, July 2017.