ON A THEORY OF NETWORK EQUIVALENCE

Ralf Koetter
ralf.koetter@tum.de

Michelle Effros
effros@caltech.edu

Muriel Médard
medard@mit.edu

***Abstract*—We describe an equivalence result for network capacity. Roughly, our main result is as follows. Given a network of noisy, independent, memoryless links, a collection of demands can be met on the given network if and only if it can be met on another network where each noisy link is replaced by a noiseless bit pipe with throughput equal to the noisy link capacity. This result was previously known only for multicast connections.**

I. Introduction

Network information theory has two natural facets reflecting different approaches to networking. On the one hand, networks are considered in the graph theoretic setup consisting of nodes and links. Typical concepts are information flows and routing issues. The links connecting nodes in a network are usually noise-free bit pipes, which can be used error free up to a certain capacity. On the other hand, multiterminal information theory deals with noisy channels, or rather the stochastic nature of input and output variables at nodes in a network. Here the interplay of transmissions in a network leads to a different set of questions dealing with fundamental limits of communication. The capacity regions of broadcast and multiple access channels are examples of questions addressed in the context of multiterminal information theory and that appear to have no obvious equivalent in networks consisting of error free bit pipes. The two views of networking are natural facets of the same problem. The objective of this paper is to explore the relationship between these two worlds.

Establishing viable bridges between these two areas shows to be surprisingly fertile. For example, questions about feedback in multiterminal systems are nicely expressed in networks of error free links. Separation issues — in particular separation between network coding and channel coding – have natural answers, revealing many network problems as combinatorial rather than statistical, even in noisy networks. In fact, many problems in network information theory appear to be reducible to solving a central network coding problem described as follows: Given a network of *error free* rate-constrained bit pipes, is a given set of demands satisfieable or not. In certain situations, most notably a multicast demand, this question has nice and simple answers. However, the general case is wide open. In fact, it is suspected that the central combinatorial network coding problem is hard, however NP hardness is only established for the class of linear network coding [1]. Nevertheless, there are algorithms available that, with running time that is exponential in the number of nodes, solve just this problem [2]. This possibility to, in principle, characterize the rate region of a combinatorial network coding problem will be a corner stone for our investigations.

It thus appears that any research into network information theory must acknowledge the fact that fully characterizing the combinatorial network coding problem is out of reach [3]. Nevertheless, moderate size networks can be solved quite efficiently. The situation is not unlike issues in complexity theory, where a lot of research is devoted to showing that one problem is essentially as difficult as another one without being able to give precise expressions as to how difficult a problem is in absolute terms. Inspired by this analogy, we resort in this paper to characterizing the relationship of a number of network coding problems to the central combinatorial network coding problem. This characterization is all we need if we want to address separation issues in networks, but also other questions, such as a degree-of-freedom or high signal to noise ratio analysis, reveal interesting insights.

It is interesting to note that the reduction of a network information theoretic question to its combinatorial essence is also at the heart of some related recent publications, see e.g. [4]. While our approach is quite different in terms of technique and also results, we believe it to be no coincidence that in both cases the reduction of a problem to its combinatorial essence is a central step.

II. Intuition and Summary of Results

To clarify the benefits of the proposed approach, consider the problem of finding the capacity region of a network of independent noisy channels. When there are general demands on such a network, i.e. any two nodes in the network may want to exchange information, this problem seems to be completely out of reach. Yet, for special demands on the network, the situation is different. For example, the case of a single unicast demand in the noise free network was solved by Ford and Fulkerson [5]. The case of a single multicast demand was later solved for noise-free networks [6] and noisy links [7], [8]. The latter papers show a separation between the combinatorial

978-1-4244-4536-3/09/$25.00 ©2009 IEEE

and statistical problems by first finding an outer bound for the desired rate region and then showing that the outer bound is reachable using channel coding to replace noisy links by lossless bit pipes and then network coding across those bit pipes. That approach cannot be used to extend this result to arbitrary demands since the cut-set outer bound is not tight and finding the true outer bound for each new demand structure is intractable.

The separated strategy is always possible, but not always optimal. For example, consider a pair of identical, independent channels each of capacity C running in parallel from node s to node r. For any $R < C$, a separated code obtains reliable communication across each link using a channel code with 2^{nR} codewords of blocklength n. Consider instead operating each channel at rate $2R > C$ by designing a single code with 2^{2nR} codewords of blocklength $2n$ and sending half of each codeword across each channel. Since decoding together achieves reliable communication and doubling the blocklength improves the error exponent, joint channel and network coding yields better error performance; it remains to be shown whether joint coding can also achieve better rates.

Roughly, our main result is as follows. A collection of demands can be met on a network of noisy, independent, memoryless links if and only if it can be met on another network where each noisy link is replaced by a noiseless bit pipe with throughput equal to the noisy link capacity.

This claim has a number of surprisingly powerful consequences. It shows separation between channel and network coding for arbitrary networks and arbitrary demands Also, many network information theoretic questions are naturally asked in the light of this combinatorial perspective. For example, it provides an alternative proof to the classical result that feedback does not increase the capacity of a point-to-point link since the min-cut between transmitter and receiver is the same with or without feedback. Most importantly, it reveals that at the heart of information theory lie combinatorial problems involving finding the rate region for error-free networks.

Since the prior approach of finding the outer bound and proving it achievable is out of the question, we build an equivalence theorem instead. Intuitively, we prove equivalence between networks A and B by showing that if anyone shows us a way to operate network A at one rate point, then we can find a way to operate network B at the same rate point and vice versa. Note that this never answers the question of whether a particular rate point is in the rate region or not. Operating codes designed for bit pipes in the noisy channel network is straight forward using a separated strategy. The other direction is harder since a noisy network allows a far richer algorithmic behavior. It is known that a noiseless bit-pipe of a given throughput can emulate any discrete memoryless channel of lesser capacity [9], so a network of bit pipes may be operated as if it were a network of noisy links. Yet applying this result seems to be difficult. Difficulties arise with continuous random variables, timing questions, and proving continuity of rate regions. Worst of all, since we do not know which strategy achieves the network capacity, we must be able to emulate all of them. We prove our main claim directly, without exploiting [9].

III. The Setup

A multiterminal network comprises m nodes. We denote the random variables transmitted and received at node i at time-step t by $X_t^{(i)} \in \mathcal{X}^{(i)}$ and $Y_t^{(i)} \in \mathcal{Y}^{(i)}$. The network is assumed to be memoryless, so it is characterized by a conditional probability distribution $p(\mathbf{y}|\mathbf{x}) = p(y^{(1)}, \ldots, y^{(m)}|x^{(1)}, \ldots, x^{(m)})$. A blocklength-$n$ code operates the network over n time steps with the goal of communicating, for all (i,j), message $W^{(i\to j)} \in \mathcal{W}^{(i\to j)} \stackrel{\text{def}}{=} \{1, \ldots, 2^{nR^{(i\to j)}}\}$ from node i to node j. The vector of rates $R^{(i\to j)}$ is denoted by $\mathcal{R}$. A network is thus written as a triple $\left(\prod_{i=1}^m \mathcal{X}^{(i)}, p(\mathbf{y}|\mathbf{x}), \prod_{i=1}^m \mathcal{Y}^{(i)}\right)$ with the added constraint that $X_t^{(i)}$ is a function of $\{Y_1^{(i)}, \ldots, Y_{t-1}^{(i)}, W^{(i\to 1)}, \ldots, W^{(i\to m)}\}$ alone.

The structure of a network is given as a directed graph G with node set $V = \{1, \ldots, m\}$ and edge set E. This paper treats point-to-point transmissions, where each edge $e \in E$ takes the form $e = [v_1, v_2]$, $v_1, v_2 \in V$. In future work, we replace edges by hyperedges to treat broadcast, multiple access, and interference transmissions.

The indegree and outdegree of node i are $d_{\text{in}}(i) \stackrel{\text{def}}{=} |\{v_1|[v_1, i] \in E\}|$ and $d_{\text{out}}(i) \stackrel{\text{def}}{=} |\{v_2|[i, v_2] \in E\}|$. When $d_{\text{out}}(i)$ and $d_{\text{in}}(i)$ exceed 1, $\mathcal{X}^{(i)} = \prod_{d=1}^{d_{\text{out}}(i)} \mathcal{X}^{(i,d)}$, $\mathcal{Y}^{(i)} = \prod_{d=1}^{d_{\text{in}}(i)} \mathcal{Y}^{(i,d)}$, $X_t^{(i)} = (X_t^{(i,1)}, \ldots, X_t^{(i,d_{\text{out}}(i))})$, and $Y_t^{(i)} = (Y_t^{(i,1)}, \ldots, Y_t^{(i,d_{\text{in}}(i))})$. We use $V_1(e)$ and $V_2(e)$ to denote the input and output ports of edge e. For example, if channel $e = [i,j]$ has input $X^{(i,s)}$ and output $Y^{(j,r)}$ for some $s \in \{1, \ldots, d_{\text{out}}(i)\}$ and $r \in \{1, \ldots, d_{\text{in}}(j)\}$, then $V_1(e) = (i,s)$ and $V_2(e) = (j,r)$.

When the network characterization corresponds to $G = (V,E)$, we factor $p(\mathbf{y}|\mathbf{x})$ as

$$\left(\prod_{i=1}^m \mathcal{X}^{(i)}, \prod_{e\in E} p(Y^{(V_2(e))}|X^{(V_1(e))}), \prod_{i=1}^m \mathcal{Y}^{(i)}\right),$$

again with the constraint that $X_t^{(i)}$ is a function of $\{Y_1^{(i)}, \ldots, Y_{t-1}^{(i)}, W^{(i\to 1)}, \ldots, W^{(i\to m)}\}$ alone. We here investigate some information theoretic aspects of replacing factors in the factorization of $p(\mathbf{y}|\mathbf{x})$.

Remark 1 In some situations it is important to be able to embed the transmissions of various nodes in a schedule which may or may not depend on the messages to be sent and the symbols that were received in the network. It is straightforward to model such a situation in the above setup by including in the input and output alphabets symbols for the case when nothing was sent on a particular link. In this way we can assume that at each time t random variables $X_t^{(i)}$ and $Y_t^{(i)}$ are given.

Definition 1 *Let a network*

$$\mathcal{N} \stackrel{\text{def}}{=} \left(\prod_{i=1}^{m} \mathcal{X}^{(i)}, \prod_{e \in E} p(y^{(V_2(e))}|x^{(V_1(e))}), \prod_{i=1}^{m} \mathcal{Y}^{(i)} \right)$$

be given corresponding to a graph $G = (V, E)$. A blocklength-n solution $\mathcal{S}(\mathcal{N})$ to this network is defined as a set of encoding and decoding functions:

$$X_t^{(i)} : (\mathcal{Y}^{(i)})^{t-1} \times \prod_{j=1}^{m} \mathcal{W}^{(i\to j)} \to \mathcal{X}^{(i)}$$
$$\hat{W}^{(j\to i)} : (\mathcal{Y}^{(i)})^{n} \times \prod_{j=1}^{m} \mathcal{W}^{(i\to j)} \to \mathcal{W}^{(j\to i)}$$

mapping $(Y_1^{(i)}, \ldots, Y_{t-1}^{(i)}, W^{(i\to 1)}, \ldots, W^{(i\to m)})$ to $X_t^{(i)}$ and $(Y_1^{(i)}, \ldots, Y_n^{(i)}, W^{(i\to 1)}, \ldots, W^{(i\to m)})$ to $\hat{W}^{(j\to i)}$. Solution $\mathcal{S}(\mathcal{N})$ is called a $(\lambda, \mathcal{R})$-solution, denoted $(\lambda, \mathcal{R}) - \mathcal{S}(\mathcal{N})$, if the encoding and decoding functions imply $\Pr(W^{(i\to j)} \neq \hat{W}^{(i\to j)}) < \lambda$ for all i, j.

Definition 2 *The rate region $\mathscr{R}(\mathcal{N}) \subset \mathbb{R}_+^{m(m-1)}$ of a network $\mathcal{N}$ is the closure of the set of rate vectors $\mathcal{R}$ such that for any $\lambda > 0$, a $(\lambda, \mathcal{R}) - \mathcal{S}(\mathcal{N})$ solution exists.*[1]

IV. Results

The goal of this paper is not to give the capacity regions of networks with respect to various demands; that is an intractable problem due to its combinatorial nature. Rather we wish to develop equivalence relationships between capacity regions of networks. Given the existence of a $(\lambda, \mathcal{R})$-$\mathcal{S}(\mathcal{N})$ solution for a network $\mathcal{N}$, we wish to imply statements for the existence of a $(\lambda', \mathcal{R}')$-$\mathcal{S}(\mathcal{N}')$ solution for a network $\mathcal{N}'$.

Assume, for example, a network contains an edge $\bar{e} = [i, j]$. The input and output random variables are $X^{(V_1(\bar{e}))} = X^{(i,s)}$ and $Y^{(V_2(\bar{e}))} = Y^{(j,r)}$. The transition probability for the network thus factors as:

$$p(y^{(j,r)}|x^{(i,s)}) \prod_{e \in E \setminus \{\bar{e}\}} p(y^{(V_2(e))}|x^{(V_1(e))}). \qquad (1)$$

[1]The blocklength n may vary with λ, $\mathcal{R}$, and $\mathcal{N}$.

Let another network $\mathcal{N}'$ be given with random variables $(\tilde{X}^{(i,s)}, \tilde{Y}^{(j,r)})$ replacing $(X^{(i,s)}, Y^{(j,r)})$ in $\mathcal{N}$. We have replaced the link characterized by $p(y^{(j,r)}|x^{(i,s)})$ with another link characterized by $\tilde{p}(\tilde{y}^{(j,r)}|\tilde{x}^{(i,s)})$. When $I(Y^{(j,r)}; X^{(i,s)}) \leqslant I(\tilde{Y}^{(j,r)}; \tilde{X}^{(i,s)})$, we want to prove that the existence of a $(\lambda, \mathcal{R})$-$\mathcal{S}(\mathcal{N})$ solution implies the existence of a $(\lambda', \mathcal{R}')$-$\mathcal{S}(\mathcal{N}')$ solution, where λ' can be made arbitrarily small if λ can. Since node j need not decode $Y^{(j,r)}$, channel capacity is not necessarily a relevant characterization of the link's behavior. For example a Gaussian channel from i to j might contribute a real-valued estimation of the input random variable; a binary erasure channel that replaces it cannot immediately deliver the same functionality.

Our proof does not invent a coding scheme. Instead, we demonstrate a technique for operating any coding scheme for $\mathcal{N}$ on the network $\mathcal{N}'$. Since there exists a coding scheme for $\mathcal{N}$ that achieves any point in the interior of $\mathscr{R}(\mathcal{N})$, this proves that $\mathscr{R}(\mathcal{N}) \subseteq \mathscr{R}(\mathcal{N}')$. Since we don't know the form of an optimal code for $\mathcal{N}$, our method must work for all possible codes on $\mathcal{N}$. For example, it must succeed even when the code for $\mathcal{N}$ is time-varying. As a result, we cannot apply typicality arguments across time. We introduce instead a notion of stacking in order to exploit averaging arguments across multiple uses of the same network not across time.

The N-fold stacked network $\underline{\mathcal{N}}$ is the network $\mathcal{N}$ repeated N times. Thus $\underline{\mathcal{N}}$ has N copies of each vertex $v \in V$ and N copies of each edge $e \in E$. Since multisets (not sets) describe the vertices and edges of $\underline{\mathcal{N}}$, the stacked network is not a network and new definitions are required. We carry over notation from $\mathcal{N}$ to $\underline{\mathcal{N}}$ using underlines. So $\underline{W^{(i\to j)}} \in \underline{\mathcal{W}^{(i\to j)}} \stackrel{\text{def}}{=} (\mathcal{W}^{(i\to j)})^N$ is the vector of messages from i to j and $\underline{X^{(i)}}_t \in \underline{\mathcal{X}^{(i)}} \stackrel{\text{def}}{=} (\mathcal{X}^{(i)})^N$ and $\underline{Y^{(i)}}_t \in \underline{\mathcal{Y}^{(i)}} \stackrel{\text{def}}{=} (\mathcal{Y}^{(i)})^N$ are the time-t vectors of channel inputs and outputs at node i. We use argument ℓ to denote the ℓ-th layer of the stack; for example $\underline{W^{(i\to j)}}(\ell)$ is the message from i to j in layer ℓ. Since $\underline{\mathcal{W}^{(i\to j)}} \stackrel{\text{def}}{=} (\mathcal{W}^{(i\to j)})^N$, the rate $R^{(i\to j)}$ for a stacked network is defined as $(\log |\underline{\mathcal{W}^{(i\to j)}}|)/(nN)$. This makes rates comparable between $\mathcal{N}$ and $\underline{\mathcal{N}}$.

Definition 3 *Let $\underline{\mathcal{N}}$ be the N-fold stacked network for a network $\mathcal{N}$. A blocklength-n solution $\mathcal{S}(\underline{\mathcal{N}})$ to $\underline{\mathcal{N}}$ is a set of encoding and decoding functions:*

$$\underline{X^{(i)}}_t : (\underline{\mathcal{Y}^{(i)}})^{t-1} \times \prod_{j=1}^{m} \underline{\mathcal{W}^{(i\to j)}} \to \underline{\mathcal{X}^{(i)}}$$
$$\underline{\hat{W}^{(j\to i)}} : (\underline{\mathcal{Y}^{(i)}})^{n} \times \prod_{j=1}^{m} \underline{\mathcal{W}^{(i\to j)}} \to \underline{\mathcal{W}^{(j\to i)}}.$$

We call $\mathcal{S}(\underline{\mathcal{N}})$ a $(\lambda, \mathcal{R})$-solution, denoted $(\lambda, \mathcal{R}) - \mathcal{S}(\underline{\mathcal{N}})$, if $\Pr(\underline{W^{(i\to j)}} \neq \underline{\hat{W}^{(i\to j)}}) < \lambda$ for all i, j.

Definition 4 *The rate region $\mathscr{R}(\underline{\mathcal{N}}) \subset \mathbb{R}_+^{m(m-1)}$ of $\underline{\mathcal{N}}$ is the closure of all rate vectors $\mathcal{R}$ such that a $(\lambda, \mathcal{R})-\mathcal{S}(\underline{\mathcal{N}})$ solution exists for any $\lambda > 0$ and all N sufficiently large.*

Theorem 1 *Rate regions $\mathscr{R}(\mathcal{N})$ and $\mathscr{R}(\underline{\mathcal{N}})$ are identical, and for each $\mathcal{R} \in$ int$(\mathscr{R}(\underline{\mathcal{N}}))$ there exist $(2^{-N\delta}, \mathcal{R}) - \mathcal{S}(\underline{\mathcal{N}})$ solutions that channel code the messages and then apply the same solution independently in each layer of $\underline{\mathcal{N}}$.*

Proof. $\mathscr{R}(\underline{\mathcal{N}}) \subseteq \mathscr{R}(\mathcal{N})$*: For any $\mathcal{R} \in$ int$(\mathscr{R}(\underline{\mathcal{N}}))$ and $\lambda > 0$, there exists a $(\lambda, \mathcal{R}) - \mathcal{S}(\underline{\mathcal{N}})$ solution to $\underline{\mathcal{N}}$. Let n be its blocklength. We build a blocklength nN solution $\mathcal{S}(\mathcal{N})$ for $\mathcal{N}$ by implementing the N layers of time step t of $\mathcal{S}(\underline{\mathcal{N}})$ in times $(t-1)N+1, \ldots, tN$ in $\mathcal{N}$. Since $\mathcal{S}(\mathcal{N})$ satisfies the causality constraints and operates the mappings from $\mathcal{S}(\underline{\mathcal{N}})$ on $\mathcal{N}$, $\mathcal{S}(\mathcal{N})$ achieves the same rate and error probability on $\mathcal{N}$ as $\mathcal{S}(\underline{\mathcal{N}})$ achieves on $\underline{\mathcal{N}}$.*

$\mathscr{R}(\underline{\mathcal{N}}) \supseteq \mathscr{R}(\mathcal{N})$*: Let $\mathcal{R} \in$ int$(\mathscr{R}(\mathcal{N}))$, and fix some $\tilde{\mathcal{R}} \in$ int$(\mathscr{R}(\mathcal{N}))$ for which $\tilde{R}^{(i\to j)} > R^{(i\to j)}$ for all i,j. Set $\rho = \min_{i,j}(\tilde{R}^{(i\to j)} - R^{(i\to j)})$. For reasons that will become clear later, we seek a solution with error probability λ and blocklength n satisfying $\max_{i,j}\tilde{R}^{(i\to j)}\lambda + h(\lambda)/n < \rho$. Such a solution is guaranteed to exist since $\tilde{\mathcal{R}} \in$ int$(\mathscr{R}(\mathcal{N}))$ implies that there exists a $(\lambda, \tilde{\mathcal{R}}) - \mathcal{S}(\mathcal{N})$ solution for any $\lambda > 0$. Fix a $(\lambda, \tilde{\mathcal{R}}) - \mathcal{S}(\mathcal{N})$ solution. To avoid confusion between the target rate $\mathcal{R}$ and the solution's higher rate $\tilde{\mathcal{R}}$, let $W^{(i\to j)}$ and $\hat{W}^{(i\to j)}$ denote the messages and their reconstructions at rate $\tilde{R}^{(i\to j)}$. The messages and reconstructions at the target rate are $U^{(i\to j)}$ and $\hat{U}^{(i\to j)}$. Let code $\mathcal{S}(\underline{\mathcal{N}})$ be the solution achieved by applying the $(\lambda, \tilde{\mathcal{R}}) - \mathcal{S}(\mathcal{N})$ solution independently in each layer of $\underline{\mathcal{N}}$. Then $(\underline{W^{(i\to j)}}(\ell), \underline{\hat{W}^{(i\to j)}}(\ell))$, $\ell \in \{1, \ldots, N\}$, constitutes N uses of a discrete memoryless channel from i to j. When the input is uniformly distributed, the mutual information between the channel input and the channel output satisfies $I(\underline{W^{(i\to j)}}(\ell); \underline{\hat{W}^{(i\to j)}}(\ell)) > n\tilde{R}^{(i\to j)} - (\lambda n\tilde{R}^{(i\to j)} + h(\lambda))$ by Fano's inequality. For each (i,j) we now apply a uniform, random $(2^{N(nR^{(i\to j)})}, N)$-channel code across the N uses of this channel. The channel code's encoder maps $\underline{U^{(i\to j)}}$ to $\underline{W^{(i\to j)}}$. Since the code's rate $nR^{(i\to j)}$ is strictly less than the channel's mutual information, precisely $I(\underline{W^{(i\to j)}}(\ell); \underline{\hat{W}^{(i\to j)}}(\ell)) - nR^{(i\to j)} > n\rho - (\lambda n\tilde{R}^{(i\to j)} + h(\lambda)) > 0$. Applying the strong coding theorem gives an error probability less than $2^{-N\delta}$, where δ is an increasing function of the gap $\min_{i,j}[I(\underline{W^{(i\to j)}}(\ell); \underline{\hat{W}^{(i\to j)}}(\ell)) - nR^{(i\to j)}]$.* ■

We henceforth restrict our attention to solutions $\mathcal{S}(\underline{\mathcal{N}})$ that first channel code messages and then apply the same solution $\mathcal{S}(\mathcal{N})$ independently in each layer of $\underline{\mathcal{N}}$. By Theorem 1, there is no loss of generality in this restriction. This and the i.i.d. uniform choice of codewords in the channel code design maintains the desired i.i.d. structure across layers. Our argument employs random coding at multiple junctures, choosing the instances only once all codes are in place. (See Theorem 3.)

We now focus on the first factor in (1), temporarily dropping superscripts (i,s) and (j,r) and referring to random variables X, Y and their realizations x, y for brevity.

Consider a $(2^{-N\delta}, \mathcal{R})$-$\mathcal{S}(\underline{\mathcal{N}})$ solution that applies the same solution $\mathcal{S}(\mathcal{N})$ in each layer of $\underline{\mathcal{N}}$. Let $p_t(x,y) = p_t(x)p(y|x)$ be the distribution imposed by $\mathcal{S}(\mathcal{N})$ across the channel $p(y|x)$ at time t. Let $A_{\epsilon,t}^{(N)}$ be the corresponding jointly typical set. Define $B_{\epsilon,t}^{(N)}(\bar{\lambda}) \stackrel{\text{def}}{=} \{(x^N, y^N) \in A_{\epsilon,t}^{(N)} : \Pr(\cup_{a,b}\{\underline{W^{(a\to b)}} \neq \underline{\hat{W}^{(a\to b)}}\}|(\underline{X}_t, \underline{Y}_t) = (x^N, y^N)) \geqslant \bar{\lambda}\}$. The probability in $B_{\epsilon,t}^{(N)}(\bar{\lambda})$ results from the operation of $\mathcal{S}(\underline{\mathcal{N}})$ on $\underline{\mathcal{N}}$.

Lemma 2 *If there exists of a $(2^{-N\delta}, \mathcal{R})$-$\mathcal{S}(\underline{\mathcal{N}})$ solution for a stacked network $\underline{\mathcal{N}}$, then for N sufficiently large*[2]

$$|B_{\epsilon,t}^{(N)}(\bar{\lambda})|/|A_{\epsilon,t}^{(N)}| < (m^2 2^{-N(\delta-2\epsilon)})/((1-\epsilon)\bar{\lambda}).$$

Proof. *Let $E = \bigcup_{a,b}\{\underline{W}^{(a\to b)} \neq \underline{\hat{W}}^{(a\to b)}\}$. Then $\Pr(B_{\epsilon,t}^{(N)}(\bar{\lambda}) \cap E) \leqslant \Pr(A_{\epsilon,t}^{(N)} \cap E) \leqslant \Pr(E) \leqslant m^2 2^{-N\delta}$ by the union bound and lemma assumption. (The assumption is valid for rates $\mathcal{R}$ of interest by Theorem 1.) Thus*

$$\begin{aligned} m^2 2^{-N\delta} &\geqslant \Pr(E|B_{\epsilon,t}^{(N)}(\bar{\lambda}))\Pr(B_{\epsilon,t}^{(N)}(\bar{\lambda})) \\ &> \bar{\lambda}|B_{\epsilon,t}^{(N)}(\bar{\lambda})|2^{-N(H(X,Y)+\epsilon)}\frac{|A_{\epsilon,t}^{(N)}|}{|A_{\epsilon,t}^{(N)}|}. \end{aligned}$$

Then $|A_{\epsilon,t}^{(N)}|2^{-N(H(X_t,Y_t)-\epsilon)} > \Pr(A_{\epsilon,t}^{(N)}) > 1-\epsilon$ for large N by the AEP gives the desired result. ■

The input $\underline{X}_t$ to channel $p(y|x)$ in $\underline{\mathcal{N}}$ usually results in a jointly typical output $\underline{Y}_t$. We argue in Theorem 3 that the functionality of $\underline{\mathcal{N}}$ remains unchanged if we replace $p(y|x)$ by any other means to pick a jointly typical $\underline{Y}_t$.

Theorem 3 *Let networks $\mathcal{N}$ and $\hat{\mathcal{N}}$ be defined as*

$$\begin{aligned} \mathcal{N} = \Big(&\mathcal{X}^{(1,1)} \times \cdots \times \mathcal{X}^{(i,s)} \times \cdots \times \mathcal{X}^{(m,d_{\text{out}}(m))}, \\ &p(y^{(j,r)}|x^{(i,s)}) \prod_{e \in E\setminus\{\bar{e}\}} p(y^{(V_2(e))}|x^{(V_1(e))}), \\ &\mathcal{Y}^{(1,1)} \times \cdots \times \mathcal{Y}^{(j,r)} \times \cdots \times \mathcal{Y}^{(m,d_{\text{in}}(m))}\Big) \end{aligned}$$

[2]We use $|A|$ to denote cardinality or volume depending on the alphabet of A. We assume entropy $H(X,Y) < \infty$ in both cases.

$$\hat{\mathcal{N}} = \Big(\mathcal{X}^{(1,1)} \times \cdots \times \hat{\mathcal{X}}^{(i,s)} \times \cdots \times \mathcal{X}^{(m,d_{\text{out}}(m))}, \delta(x^{(i,s)} - y^{(j,r)}) \prod_{e \in E\setminus\{\bar{e}\}} p(y^{(V_2(e))}|x^{(V_1(e))}), \mathcal{Y}^{(1,1)} \times \cdots \times \hat{\mathcal{Y}}^{(j,r)} \times \cdots \times \mathcal{Y}^{(m,d_{\text{in}}(m))}\Big),$$

where $(\hat{\mathcal{X}}^{(i,s)}, \delta(\hat{x}^{(i,s)} - \hat{y}^{(j,r)}), \hat{\mathcal{Y}}^{(j,r)})$ *is a bit pipe that noiselessly maps* $R > C \stackrel{\text{def}}{=} \max_{p(x^{i,s})} I(X^{(i,s)}; Y^{(j,r)})$ *bits from its input to its output at each time step. Then* $\mathscr{R}(\mathcal{N}) \subseteq \mathscr{R}(\hat{\mathcal{N}})$.

Proof. Let $\underline{\mathcal{N}}$ and $\underline{\hat{\mathcal{N}}}$ be the stacked networks for $\mathcal{N}$ and $\hat{\mathcal{N}}$. Since $\mathscr{R}(\mathcal{N}) = \mathscr{R}(\underline{\mathcal{N}})$ and $\mathscr{R}(\hat{\mathcal{N}}) = \mathscr{R}(\underline{\hat{\mathcal{N}}})$ by Theorem 1, it suffices to prove $\mathscr{R}(\underline{\mathcal{N}}) \subseteq \mathscr{R}(\underline{\hat{\mathcal{N}}})$. For any $\mathcal{R} \in int(\mathscr{R}(\underline{\mathcal{N}}))$ and any $\lambda > 0$, there exists a $(\lambda, \mathcal{R})$-$\mathcal{S}(\underline{\mathcal{N}})$ solution for network $\underline{\mathcal{N}}$. We prove that there also exists a $(\hat{\lambda}, \mathcal{R})$-$\mathcal{S}(\underline{\hat{\mathcal{N}}})$ solution for all $\hat{\lambda} > 0$.

By Theorem 1, $\mathcal{R} \in int(\mathscr{R}(\underline{\mathcal{N}}))$ *implies that there exists a* $(2^{-N\delta}, \mathcal{R})$-$\mathcal{S}(\underline{\mathcal{N}})$ *solution for* $\underline{\mathcal{N}}$ *that applies the same solution* $\mathcal{S}(\mathcal{N})$ *in every layer of* $\underline{\mathcal{N}}$. *Let* $p_t(x^N, y^N) = \prod_{\ell=1}^N p_t(x_\ell, y_\ell)$ *be the distribution of* $(\underline{X^{(i,s)}}_t, \underline{Y^{(j,r)}}_t)$ *using solution* $\mathcal{S}(\underline{\mathcal{N}})$ *on* $\underline{\mathcal{N}}$. *Let* n *be the blocklength of* $\mathcal{S}(\underline{\mathcal{N}})$. *For each* $t \in \{1, \ldots, n\}$, *randomly design a source code* $(\alpha_{N,t}, \beta_{N,t})$ *with* 2^{NR} *codewords drawn i.i.d. according to* $p_t(y^N) = \sum_{x^N \in \mathcal{X}^N} p_t(x^N, y^N)$ *and an encoder* $\alpha_{N,t}$ *that maps each* x^N *to an index* k *for which* $(x^N, \beta_{N,t}(k)) \in A_{\epsilon,t}^{(N)} \setminus B_{\epsilon,t}^{(N)}(\bar{\lambda})$ *if available.*

We now run $\mathcal{S}(\underline{\mathcal{N}})$ *on* $\underline{\hat{\mathcal{N}}}$ *by mapping* $\underline{X^{(i,s)}}_t$ *to* $\underline{\hat{X}^{(i,s)}}_t = \alpha_{N,t}(\underline{X^{(i,s)}}_t)$ *before transmission by node* i *and mapping* $\underline{\hat{Y}^{(j,r)}}_t$ *to* $\underline{Y^{(j,r)}}_t = \beta_{N,t}(\underline{\hat{Y}^{(j,r)}}_t)$ *after receipt by node* j.

We offer the following analysis for the error probability. The expectation is over all random code designs.

$$\begin{aligned} &E\Pr\Big(\cup_{(a,b)}\{\underline{W^{(a\to b)}} \neq \underline{\hat{W}^{(a\to b)}}\}\Big) \\ &\leqslant \Big[\sum_{t=1}^{n} E\Pr((\underline{X^{(i,s)}}_t, \underline{Y^{(j,r)}}_t) \notin A_{\epsilon,t}^{(N)} \setminus B_{\epsilon,t}^{(N)}(\bar{\lambda}), \\ &\quad (\underline{X^{(i,s)}}_{t'}, \underline{Y^{(j,r)}}_{t'}) \in A_{\epsilon,t'}^{(N)} \setminus B_{\epsilon,t'}^{(N)}(\bar{\lambda}), \forall t' < t\})\Big] \\ &\quad + E\Pr(\cup_{(a,b)}\{\underline{W^{(a\to b)}} \neq \underline{\hat{W}^{(a\to b)}}\}, \\ &\quad (\underline{X^{(i,s)}}_t, \underline{Y^{(j,r)}}_t) \in A_{\epsilon,t}^{(N)} \setminus B_{\epsilon,t}^{(N)}(\bar{\lambda}) \forall t \leqslant n). \quad (2) \end{aligned}$$

Let $K_t(x^N, y^N) = 1$ *if* $(x^N, y^N) \in A_{\epsilon,t}^{(N)} \setminus B_{\epsilon,t}^{(N)}(\bar{\lambda})$ *and* 0 *otherwise. Note that* $(1 - ab)^k \leqslant 1 - a + e^{-bk}$ *and* $p_t(y^N) > p(y^n|x^n)2^{-N(I(X_t^{(i,s)};Y_t^{(j,r)})+3\epsilon)}$ *for all* $(x^N, y^N) \in A_{\epsilon,t}^{(N)}$ *[10, Lemmas 10.5.2, 10.5.3], where* $I(X_t^{(i,s)}; Y_t^{(j,r)}) \leqslant C$ *is the mutual information for* $p_t(x, y) = p_t(x)p(y|x)$. *Then term* t *in the sum is at most*

$$\sum_{x^N \in \mathcal{X}^N} E\hat{p}_t(x^N)(1 - \sum_{y^N} p_t(y^N)K_t(x^N, y^N))^{2^{NR}}$$

$$< \sum_{x^N} \hat{p}_t(x^N)(1 - \sum_{y^N} p(y^N|x^N)K_t(x^N, y^N) + e^{-2^{N(R - I(X^{(i,s)};Y^{(j,r)}) - 3\epsilon)}}),$$

where $\hat{p}_t(x^N)$ *is the probability of* x^N *at time* t *in the given solution. The exponential goes to zero since* $R > C$. *The remaining difference goes to zero by the AEP if* $\mathcal{N}$ *has no cycles and by analysis of the probability that* $\underline{X^{(i,s)}}_t$ *is atypical when* $(\underline{X^{(i,s)}}_{t'}, \underline{Y^{(j,r)}}_{t'}) \in A_{\epsilon,t'}^{(N)} \setminus B_{\epsilon,t'}^{(N)}(\bar{\lambda})$ *for all* $t' < t$ *otherwise. We bound the last term in (2) by* $2^{6\epsilon Nn}\bar{\lambda}/(1-\epsilon)^n$ *by first bounding the distribution* $\hat{p}_t(y^N|x^N)$ *for each random source code.*

Since the expected error probability with respect to the joint distribution over all codes can be made arbitrarily small, there exists a single instance of the collection of codes that does at least as well. ■

Remark 2 By Theorem 3, when $R < C$, we can run any solution $\mathcal{S}(\underline{\mathcal{N}})$ on $\underline{\hat{\mathcal{N}}}$ by source coding across the layers; thus $\mathscr{R}(\mathcal{N}) \subseteq \mathscr{R}(\hat{\mathcal{N}})$ when $R < C$ by Theorem 1. By Shannon's channel coding theorem, when $R > C$, we can run any solution $\mathcal{S}(\underline{\hat{\mathcal{N}}})$ on $\underline{\mathcal{N}}$ by channel coding across the layers; thus $\mathscr{R}(\mathcal{N}) \supseteq \mathscr{R}(\hat{\mathcal{N}})$ when $R > C$ by Theorem 1. Together, these results give the desired equivalence. As usual, point $R = C$ remains ambiguous. Repeating this argument for each $e \in E$ relates a network of channels to a network of noiseless bit pipes.

References

[1] M. Langberg, M. Sprintson, and J. Bruck, "Network coding: a computational perspective," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 145–157, 2008.

[2] A. Subramanian and A. Thangaraj, "A simple algebraic formulation for the scalar linear network coding problem," *ArXiv e-prints*, July 2008.

[3] T. Chan and A. Grant, "Dualities between entropy function and network codes," *IEEE Transactions on Information Theory*, vol. 54, pp. 4470–4487, Oct. 2008.

[4] A. Avestimehr, S. Diggavi, and D. Tse, "Approximate capacity of Gaussian relay networks," in *Proc. of the IEEE International Symposium on Information Theory*, pp. 474–478, July 2008.

[5] L. Ford and D. Fulkerson, "Maximal flow through a network," Canadian Journal of Mathematics, vol. 8, pp. 399–404, 1956.

[6] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. IT-46, pp. 1204–1216, July 2000.

[7] S. Borade, "Network information flow: limits and achievability," in *Proceedings of the IEEE International Symposium on Information Theory*, p. 139, July 2002.

[8] L. Song, R. W. Yeung, and N. Cai, "A separation theorem for single-source network coding," *IEEE Transactions on Information Theory*, vol. 52, pp. 1861–1871, May 2006.

[9] C. Bennett, P. Shor, J. Smolin, and A.V.Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem," *IEEE Transactions on Information Theory*, vol. 48, pp. 2637–2655, Oct. 2002.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, second ed., 2006.