Wireless Network Information Flow

Amir Salman Avestimehr
Wireless Foundations
UC Berkeley,
Berkeley, California, USA.
Email: avestime@eecs.berkeley.edu

Suhas N. Diggavi
School of Computer and
Communication Sciences, EPFL,
Lausanne, Switzerland.
Email: suhas.diggavi@epfl.ch

David N C. Tse
Wireless Foundations
UC Berkeley,
Berkeley, California, USA.
Email: dtse@eecs.berkeley.edu

***Abstract*— We present an achievable rate for general deterministic relay networks, with broadcasting at the transmitters and interference at the receivers. In particular we show that if the optimizing distribution for the information-theoretic cut-set bound is a product distribution, then we have a complete characterization of the achievable rates for such networks. For linear deterministic finite-field models discussed in a companion paper [3], this is indeed the case, and we have a generalization of the celebrated max-flow min-cut theorem for such a network.**

I. Introduction

Consider a network represented by a directed relay network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ are the vertices representing the communication nodes in the relay network. The communication problem considered is unicast (or multicast with all destinations requesting the *same* message). Therefore a special node $S \in \mathcal{V}$ is considered the source of the message and a special node $D \in \mathcal{V}$ is the intended destination. All other nodes in the network facilitate communication between S and D. In a wireline network, such as studied in [1], the edges $\mathcal{E}$ of the network do not interact and are orthogonal communication channels. In this paper, transmissions are *not necessarily* orthogonal and signals sent by the nodes in $\mathcal{V}$ can in general broadcast and also interfere with one another. In particular, for each vertex $j \in \mathcal{V}$ of the network, there is only one transmitted signal x_j which is broadcast to the other nodes connected to this vertex. Moreover it has only one received signal y_j which is a deterministic function of all the signals transmitted by the nodes connected to it. By connection we mean the nodes that have edges belonging to the set $\mathcal{E}$. By deterministic we mean that $y_j = g_j(\{x_k\}_{k \in \mathcal{N}_j})$, where $\mathcal{N}_j$ is the input neighbors of node j. Therefore, we have deterministic broadcast and multiple access channels incorporated into the model to reflect physical layer effects.

This approach is motivated by the development of the linear deterministic finite-field model for wireless channels [2], and its connection to Gaussian relay networks [3]. Historically, deterministic relay networks were perhaps first studied in [4], where a deterministic model with broadcast but *no multiple access* was studied (the so-called Aref's networks). For such a network, the unicast capacity was determined in [4] and its extension to multicast capacity when all receivers needed the same message was done in [9]. A three-node deterministic relay network capacity was characterized in [10], where *both* broadcast and multiple access were allowed. Network coding is information flow on a very special class of deterministic networks, where all the links are non-interfering and orthogonal. For such networks, the unicast capacity is given by the classical max-flow min-cut theorem of Ford-Fulkerson, and the multicast capacity has been determined in the seminal work [1]. More recently, the capacity of a class of erasure relay networks has been established where random erasures attempt to model the noise and collisions [12]. In all these cases, where the characterization exists, the information-theoretic cut-set was achievable. Recently, a relay network where the cut-set bound is not tight has been demonstrated in [5].

We first consider general deterministic functions to model the broadcast and multiple access channels. For such networks we show an achievability which is tight only for functions and networks where the independent input distribution optimizes the information-theoretic cut-set bound. For Aref's networks where there is no interference, this is indeed the case and our result is a generalization of his. For deterministic networks where there *is* interference but the deterministic functions are linear over a finite field, it turns out that the cut-set bound is also optimized by the product distribution. For this case, our result is a natural generalization of the celebrated max-flow min-cut theorem. These ideas are easily extended to the multicast case, where we want to simultaneously transmit one message from S

to all destinations D in the set $\mathcal{D}$. For the linear finite-field model, we characterize the multicast capacity, and therefore generalize the result in [1]. We will discuss this in more detail in the next section.

II. Problem statement and main results

A. General Deterministic network

As stated in Section I, we consider a directed network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the received signal y_j at node $j \in \mathcal{V}$ is given by

$$y_j = g_j(\{x_i\}_{i \in \mathcal{N}_j}), \tag{1}$$

where we define the input neighbors $\mathcal{N}_j$ of j as the set of nodes whose transmissions affect j, and can be formally defined as $\mathcal{N}_j = \{i : (i,j) \in \mathcal{E}\}$. Note that this implies a deterministic multiple access channel for node j and a deterministic broadcast channel for the transmitting nodes.

For any relay network, there is a natural information-theoretic cut-set bound [6], which upperbounds the reliable transmission rate R. Applied to our model, we have:

$$\begin{aligned} R &< \max_{p(\{x_j\}_{j \in \mathcal{V}})} \min_{\Omega \in \Lambda_D} I(Y_{\Omega^c}; X_\Omega | X_{\Omega^c}) \\ &\overset{(a)}{=} \max_{p(\{x_j\}_{j \in \mathcal{V}})} \min_{\Omega \in \Lambda_D} H(Y_{\Omega^c} | X_{\Omega^c}) \end{aligned} \tag{2}$$

where $\Lambda_D = \{\Omega : S \in \Omega, D \in \Omega^c\}$ is all source-destination cuts (partitions) and (a) follows since we are dealing with deterministic networks.

The following are our main results for general deterministic networks.

Theorem 2.1: Given a general deterministic relay network (with broadcast and multiple access), we can achieve all rates R up to,

$$\max_{\prod_{i \in \mathcal{V}} p(x_i)} \min_{\Omega \in \Lambda_D} H(Y_{\Omega^c} | X_{\Omega^c}) \tag{3}$$

This theorem easily extended to the multicast case, where we want to simultaneously transmit one message from S to all destinations in the set $D \in \mathcal{D}$:

Theorem 2.2: Given a general deterministic relay network (with broadcast and multiple access), we can achieve all rates R from S multicasting to all destinations $D \in \mathcal{D}$ up to,

$$\max_{\prod_{i \in \mathcal{V}} p(x_i)} \min_{D \in \mathcal{D}} \min_{\Omega \in \Lambda_D} H(Y_{\Omega^c} | X_{\Omega^c}) \tag{4}$$

This achievability result in Theorem 2.1 extends the results in [9] where only deterministic broadcast network (with no interference) were considered.

Note that when we compare (3) to the cut-set upper bound in (2), we see that the difference is in the maximizing set *i.e.*, we are only able to achieve independent (product) distributions whereas the cut-set optimization is over any arbitrary distribution. In particular, if the network and the deterministic functions are such that the cut-set is optimized by the product distribution, then we would have matching upper and lower bounds. This indeed happens when we consider the linear finite-field model discussed below.

B. Linear Finite-Field Deterministic network

A special deterministic model which is motivated [3] by its close connection to the Gaussian model is the linear finite-field model, where the received signal $\mathbf{y}_j \in \mathbb{F}_p^q$ is a vector defined over a finite field $\mathbb{F}_p$ given by,

$$\mathbf{y}_j = \sum_{i \in \mathcal{V}} \mathbf{G}_{i,j} \mathbf{x}_i, \tag{5}$$

where the transmitting signals $\mathbf{x}_k \in \mathbb{F}_p^q$, and the "channel" matrices $\mathbf{G}_{i,j} \in \mathbb{F}_p^{q \times q}$. All the operations are done over the finite field $\mathbb{F}_p$, and the network $\mathcal{G}$, implies that $\mathbf{G}_{i,j} = \mathbf{0}, i \notin \mathcal{N}_j$ reducing the sum in (5) from $N = |\mathcal{V}|$ terms *i.e.*, all transmitting nodes in the network, to just the input neighbors of j.

If we look at the cut-set upper bound for general deterministic networks (2), it is easy to see in a special case of linear finite-field deterministic networks that all cut values are simultaneously optimized by independent and uniform distribution of $\{x_i\}_{i \in \mathcal{V}}$. Moreover the optimum value of each cut Ω is logarithm of the size of the range space of the transfer matrix $\mathbf{G}_{\Omega,\Omega^c}$ associated with that cut, *i.e.*, the matrix relating the super-vector of all the inputs at the nodes in Ω to the super-vector of all the outputs in Ω^c induced by (5). This yields the following complete characterization as the corollaries of theorem 2.1 and 2.2:

Corollary 2.3: Given a linear finite-field relay network (with broadcast and multiple access), the capacity C of such a relay network is given by,

$$C = \min_{\Omega \in \Lambda_D} \operatorname{rank}(\mathbf{G}_{\Omega,\Omega^c}) \log p. \tag{6}$$

Corollary 2.4: Given a linear finite-field relay network (with broadcast and multiple access), the multicast capacity C of such a relay network is given by,

$$C = \min_{D \in \mathcal{D}} \min_{\Omega \in \Lambda_D} \operatorname{rank}(\mathbf{G}_{\Omega,\Omega^c}) \log p. \tag{7}$$

For a single source-destination pair the result in Corollary 2.3 generalizes the classical max-flow min-cut theorem for wireline networks and for multicast, the result in Corollary 2.4 generalizes the network coding result in [1] where in both these earlier results, the communication links are orthogonal. Moreover, as we will see in the proof, the encoding functions at the relay nodes could be restricted to linear functions to obtain the result in Corollary 2.3.

C. *Proof Strategy*

Theorem 2.1 is the main result of the paper and the rest of the paper is devoted to proving it. First we focus on networks that have a layered structure, i.e. all paths from the source to the destination have equal lengths. With this special structure we get a major simplification: a sequence of messages can each be encoded into a block of symbols and the blocks do not interact with each other as they pass through the relay nodes in the network. The proof of the result for layered network is similar in style to the random coding argument in [1]. We do this in sections III, IV and V, first for the linear finite-field model in III and IV and then for the general deterministic model in V. Second, we extend the result to an arbitrary network by considering its time-expanded representation. The time-expanded network is layered and we can apply our result in the first step to it. To complete the proof of the result, we need to establish a connection between the cut values of the time-expanded network and those of the original network. We do this using sub-modularity properties of entropy in Section VI[1].

III. Linear Model: An Example

In this section we give the encoding scheme for the linear deterministic model of (5) in Section III-A. In Section III-B we illustrate the proof techniques on a simple linear unicast relay network example.

A. *Encoding for linear deterministic model*

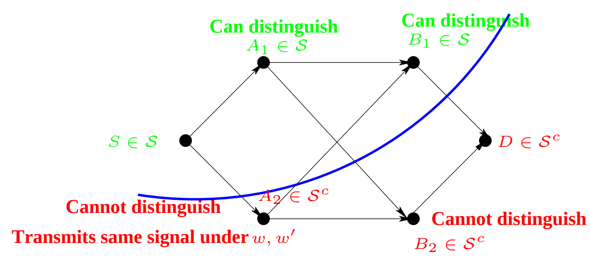
We have a single source S with message $W \in \{1, 2, \dots, 2^{TKR}\}$ which is encoded by the source S into a signal over KT transmission times (symbols), giving an overall transmission rate of R. Each relay operates over blocks of time T symbols, and uses a mapping $f_j^{(k)} : \mathcal{Y}_j^T \to \mathcal{X}_j^T$ its received symbols from the previous block of T symbols to transmit signals in the next block. In particular, block k of T received symbols is denoted by $\mathbf{y}_j^{(k)} = \{\mathbf{y}_j^{[(k-1)T+1]}, \dots, \mathbf{y}_j^{[kT]}\}$ and the transmit symbols by $\mathbf{x}_j^{(k)}$. For the model (5), we will use linear mappings $f_j(\cdot)$, *i.e.*,

$$\mathbf{x}_j^{(k)} = \mathbf{F}_j^{(k)} \mathbf{y}_j^{(k-1)}, \tag{8}$$

where $\mathbf{F}_j^{(k)}$ is chosen uniformly randomly over all matrices in $\mathbb{F}_p^{q \times q}$. Each relay does the encoding prescribed by (8). Given the knowledge of all the encoding functions $\mathbf{F}_j$ at the relays and signals received over $K + |\mathcal{V}| - 2$ blocks, the decoder $D \in \mathcal{D}$, attempts to decode the message W sent by the source.

B. *Proof illustration*

In order to illustrate the proof ideas of Theorem (2.1) we examine the network shown in Figure III-B. We will analyze this network first for linear deterministic model and then we use the same example to illustrate the ideas for general deterministic functions in Section V-B.



The network given in Figure III-B is an example of a *layered* network where the number of "hops" for each path from S to D is equal to 3 in this case[2]. The key simplification that occurs for layered networks is that we can divide the message W into K parts (sub-messages), each taking values in $w_k \in \{1, 2, \dots, 2^{TR}\}, k = 1, \dots, K$. By doing this in Figure III-B, we see that for example, nodes A_1, A_2 are sending signals which pertain to the same sub-message w_k. Therefore, the "interfering" signals in node B_1 are both about the same sub-message. This is a statement that holds in general for layered networks. For example in block number $k = 3$, the source is sending a signal about w_3, A_1, A_2 are sending signals that depend on w_2 and B_1, B_2 in turn are sending a signal to D which depends on w_1. This message synchronization implies that we can focus our attention on the error probability of a single sub-message $w = w_1$ without loss of generality.

[1]The concept of time-expanded representation is also used in [1], but the use there is to handle cycles. Our main use is to handle interaction between messages transmitted at different times, an issue that only arises when there is interference at nodes.

[2]Note that in the equal path network we do not have "self-interference" since all path-lengths from S to D in terms of "hops" are equal, though as we will see in the analysis that can easily be taken care of. However we do allow for self-interference in the model and we choose to handle such loops, and more generally cyclic networks, through time-expansion as will be seen in Section VI.

Now, since we have a deterministic network, the message w will be mistaken for another message w' is if the received signal $\mathbf{y}_D^{(3)}(w)$ under w, is the same as that would have been received under w'. This leads to a notion of *distinguishability*, which is that messages w, w' are distinguishable at any node j if $\mathbf{y}_j(w) \neq \mathbf{y}_j(w')$.

The probability of error at decoder D can be upper bounded using the union bound as,

$$P_e \leq 2^{RT}\mathbb{P}\left\{w \to w'\right\} = 2^{RT}\mathbb{P}\left\{\mathbf{y}_D^{(3)}(w) = \mathbf{y}_D^{(3)}(w')\right\}. \tag{9}$$

For the deterministic network, this event, is random only due to the randomness in the encoder map. Therefore, the probability of this event depends on the probability that we choose such an encoder map. Now, we can write,

$$\mathbb{P}\left\{w \to w'\right\} = \sum_{\Omega \in \Lambda_D} \underbrace{\mathbb{P}\left\{\text{Nodes in } \Omega \text{ can distinguish } w, w' \text{ and nodes in } \Omega^c \text{ cannot}\right\}}_{\mathcal{P}} \tag{10}$$

since the events that correspond to occurrence of the distinguishability sets $\Omega \in \Lambda_D$ are disjoint. Let us examine one term in the summation in (10). The distinguishability of $w = w_1$ from $w' = w_1'$ for the nodes A_1, A_2 are from signals $\mathbf{y}_{A_1}^{(1)}, \mathbf{y}_{A_2}^{(1)}$, for the nodes B_1, B_2 are from signals $\mathbf{y}_{B_1}^{(2)}, \mathbf{y}_{B_2}^{(2)}$ and for the receiver D it is $\mathbf{y}_D^{(3)}(w)$. For notational simplicity we will drop the block numbers associated with the transmitted and received signals for this analysis.

For the cut $\Omega = \{S, A_1, B_1\}$, a necessary condition for the distinguishability set to be this cut is that $\mathbf{y}_{A_2}(w) = \mathbf{y}_{A_2}(w')$, along with $\mathbf{y}_{B_2}(w) = \mathbf{y}_{B_2}(w')$ and $\mathbf{y}_D(w) = \mathbf{y}_D(w')$. Since the source does a random linear mapping of the message onto $\mathbf{x}_S(w)$, the probability that $\mathbf{y}_{A_2}(w) = \mathbf{y}_{A_2}(w')$ is given by,

$$\mathbb{P}\left\{(\mathbf{I}_T \otimes \mathbf{G}_{S,A_2})(\mathbf{x}_S(w) - \mathbf{x}_S(w')) = \mathbf{0}\right\} = p^{-T\text{rank}(\mathbf{G}_{S,A_2})}, \tag{11}$$

since the random mapping given in (8) induces independent uniformly distributed $\mathbf{x}_S(w), \mathbf{x}_S(w')$. Here, $\otimes$ is the Kronecker matrix product. Now, in order to analyze the probability that $\mathbf{y}_{B_2}(w) = \mathbf{y}_{B_2}(w')$, we see that since $\mathbf{y}_{A_2}(w) = \mathbf{y}_{A_2}(w')$, $\mathbf{x}_{A_2}(w) = \mathbf{x}_{A_2}(w')$, *i.e.*, the *same* signal is sent under both w, w'. Therefore, we get the probability of $\mathbf{y}_{B_2}(w) = \mathbf{y}_{B_2}(w')$ given that the distinguishability set is $\Omega = \{S, A_1, B_1\}$, as,

$$\mathbb{P}\left\{(\mathbf{I}_T \otimes \mathbf{G}_{A_1,B_2})(\mathbf{x}_{A_1}(w) - \mathbf{x}_{A_1}(w')) = \mathbf{0}\right\} = p^{-T\text{rank}(\mathbf{G}_{A_1,B_2})}. \tag{12}$$

Similarly we get,

$$\begin{aligned}
&\mathbb{P}\left\{\mathbf{y}_D(w) = \mathbf{y}_D(w') | \text{distinguishability set } \Omega\right\} \\
&\quad = \mathbb{P}\left\{(\mathbf{I}_T \otimes \mathbf{G}_{B_1,D})(\mathbf{x}_{B_1}(w) - \mathbf{x}_{B_1}(w')) = \mathbf{0}\right\} \\
&\quad = p^{-T\text{rank}(\mathbf{G}_{B_1,D})}.
\end{aligned} \tag{13}$$

Putting these together, since all three would need to occur, we see that in (10), for the network in Figure III-B, we have,

$$\begin{aligned}
\mathcal{P} &\leq p^{-T\text{rank}(\mathbf{G}_{S,A_2})}p^{-T\text{rank}(\mathbf{G}_{A_1,B_2})}p^{-T\text{rank}(\mathbf{G}_{B_1,D})} \\
&= p^{-T\{\text{rank}(\mathbf{G}_{S,A_2})+\text{rank}(\mathbf{G}_{A_1,B_2})+\text{rank}(\mathbf{G}_{B_1,D})\}}
\end{aligned} \tag{14}$$

Note that since in this example,

$$\mathbf{G}_{\Omega,\Omega^c} = \begin{bmatrix} \mathbf{G}_{S,A_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{A_1,B_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_{B_1,D} \end{bmatrix},$$

the upper bound for $\mathcal{P}$ in (14) is exactly $2^{-T\text{rank}(\mathbf{G}_{\Omega,\Omega^c})}$. Therefore, by substituting this back into (10) and (9), we see that

$$P_e \leq 2^{RT}|\Lambda_D|p^{-T\min_{\Omega\in\Lambda_D}\text{rank}(\mathbf{G}_{\Omega,\Omega^c})}, \tag{15}$$

which can be made as small as desired if $R < \min_{\Omega\in\Lambda_D}\text{rank}(\mathbf{G}_{\Omega,\Omega^c})\log p$, which is the result claimed in Corollory 2.3.

These ideas motivate first focussing on layered networks as done in Section IV. The major simplification that we get in this case is that the signals associated with different messages do not get mixed in the network and hence we can only focus on one message. Note that another simplification in layered (equal path) networks is that for a given node j, it is enough to choose the same encoding function f_j for each block k.

Now the general result for layered networks are proved in two parts: first for linear deterministic model and then for general deterministic model.

IV. Layered networks: linear deterministic model

In this section we prove main corollaries 2.3 and 2.4 for layered networks. In a layered network, for each node j we have a length l_j from the source and all the incoming signals to node j are from nodes i whose distance from the source are $l_i = l_j - 1$. Therefore, as in the example network of Figure III-B, we see that there is message synchronization, *i.e.*, all signals arriving at node j are encoding the same sub-message.

Suppose message w_k is sent by the source in block k, then since each relay j operates only on block of lengths T, the signals received at block k at any relay pertain to only message w_{k-l_j} where l_j is the path length from source to relay j. To explicitly indicate this we denote by $\mathbf{y}_j^{(k)}(w_{k-l_j}) \in \mathbb{F}_p^{qT}$ as the received signal at block k at node j. We also denote the transmitted signal at block k

as $\mathbf{x}_j^{(k)}(w_{k-1-l_j}) \in \mathbb{F}_p^{qT}$ which is obtained by randomly mapping $\mathbf{y}_j^{(k-1)}(w_{k-1-l_j}) \in \mathbb{F}_p^{qT}$.

Since we have a layered network, without loss of generality consider the message $w = w_1$ transmitted by the source at block $k = 1$. At node j the signals pertaining to this message are received by the relays at block l_j. We analyze a l_D-layer network, each layer is a MIMO sub-network. Therefore, as in the analysis of (10), we see that

$$P_e^{(D)} \leq 2^{RT} \sum_{\Omega \in \Lambda_D} \underbrace{\mathbb{P}\left\{\text{Nodes in } \Omega \text{ can distinguish } w, w' \text{ and nodes in } \Omega^c \text{ cannot}\right\}}_{\mathcal{P}} \quad (16)$$

We define $\mathbf{G}_{\Omega,\Omega^c}$ as the transfer matrix associated with the nodes in Ω to the nodes in Ω^c. Note that since we have a layered network this transfer matrix breaks up into block diagonal elements corresponding to each of the l_D layers of the network. More precisely, we can create $d = l_D$ disjoint sub-networks of nodes corresponding to each layer of the network, with $\beta_l(\Omega)$ nodes at distance $l-1$ from S that are in Ω, on one side and $\gamma_l(\Omega)$ nodes at distance l from S that are in Ω^c, on the other, for $l = 1, \ldots, l_D$.

Each node $i \in \beta_l(\Omega)$ sees a signal related to $w = w_1$ in block $l_i = l-1$, and therefore waits to receive this block and then does a random mapping to $\mathbf{x}_i^{(l_i)}(w) \in \mathbb{F}_p^{qT}$ The random mapping is done as in (8), by choosing a random matrix $\mathbf{F}_i$ of size $Tq \times Tq$ and creating

$$\mathbf{x}_i^{(l_i)}(w) = \mathbf{F}_i \mathbf{y}_i^{(l_i-1)}(w) \quad (17)$$

The received signals in the nodes $j \in \gamma_l(\Omega)$ are linear transformations of the transmitted signals from nodes $\mathcal{T}_l = \{u : (u,v) \in \mathcal{E}, v \in \gamma_l(\Omega)\}$. That is, its output depends not only on the transmitters in β_l, but also other transmitters at distance $l-1$ from S that are part of Ω^c. Since all the receivers in γ_l are at distance l from S, they form the receivers of the MIMO layer l, and we denote this vector received signal as $\mathbf{z}_l(w)$, and this can be done for all layers $l = 1, \ldots, l_D$. Note that as in the example network of Section III-B, for all the transmitting nodes in $\mathcal{T}$ which cannot distinguish between w, w' the transmitted signal would be the same under both w and w'. Therefore, in order to calculate the probability that nodes in γ_l cannot distinguish between w, w' or that $\mathbf{z}_l(w) - \mathbf{z}_l(w') = \mathbf{0}$, we see that

$$\mathbf{z}_l(w) - \mathbf{z}_l(w') = \tilde{\mathbf{G}}_l \left[\mathbf{u}_l(w) - \mathbf{u}_l(w')\right], \ l = 1, \ldots, d \quad (18)$$

where the transmitted signals from $\beta_1, \ldots, \beta_d$ are clubbed together[3] and denoted by $\mathbf{u}_l(w), l = 1, \ldots, d$. Also, due to the time-invariant channel conditions we see that $\tilde{\mathbf{G}}_l = \mathbf{I}_T \otimes \mathbf{G}_l$, where $\otimes$ is the Kronecker product. Since we are trying to calculate the probability that $\mathbf{z}_l(w) = \mathbf{z}_l(w'), l = 1, \ldots, d$, and hence we need to find the probability that $\mathbf{u}_l(w) - \mathbf{u}_l(w')$ lies in the null space of $\mathbf{G}_l$ for *each* $l = 1, \ldots, d$.

Now, if the distinct signals $\mathbf{y}_i^{(l_i)}(w), \mathbf{y}_i^{(l_i)}(w')$ received at the nodes $i \in \beta_l$ could be *jointly uniformly and independently* mapped to the transmitted signals $\mathbf{u}_l(w), \mathbf{u}_l(w')$, then we could say that the probability of this occurrence is $\frac{\text{size of null space}}{\text{size of whole space}}$. Clearly this is given by,

$$\mathbb{P}\left\{\mathbf{u}_l(w) - \mathbf{u}_l(w') \in \mathcal{N}(\tilde{\mathbf{G}}_l)\right\} = p^{-\mathrm{rank}(\tilde{\mathbf{G}}_l)} = p^{-T\mathrm{rank}(\mathbf{G}_l)}. \quad (19)$$

However, even though the signals $\mathbf{y}_i^{(l_i)}(w)$ are uniformly randomly mapped *individually* at each node $i \in \beta_l$, the overall map across all nodes in β_l is also uniform, and hence the probability given in (19) is the correct one. Since the events in each of the stages/clusters are independent, we get that

$$\mathbb{P}\left\{\mathbf{u}_l(w) - \mathbf{u}_l(w') \in \mathcal{N}(\tilde{\mathbf{G}}_l), l = 1, \ldots, d\right\} = \prod_{l=1}^{d} p^{-\mathrm{rank}(\tilde{\mathbf{G}}_l)}$$
$$= p^{-T\sum_{l=1}^{d} \mathrm{rank}(\mathbf{G}_l)}$$

Therefore, we see that

$$\mathcal{P} \leq p^{-T\sum_{l=1}^{d} \mathrm{rank}(\mathbf{G}_l)}. \quad (20)$$

Now the probability of mistaking w for w' at receiver $D \in \mathcal{D}$ is therefore

$$\begin{aligned} \mathbb{P}\left\{w \to w'\right\} &\leq \sum_{\Omega \in \Lambda_D} p^{-T\sum_{l=1}^{d(\Omega)} \mathrm{rank}(\mathbf{G}_l(\Omega))} \\ &\leq 2^{|\mathcal{V}|} p^{-T\min_{\Omega \in \Lambda} \mathrm{rank}(\mathbf{G}_{\Omega,\Omega^c})}, \end{aligned}$$

where we have used $|\Lambda_D| \leq 2^{|\mathcal{V}|}$. Note that we have used the fact that since $\mathbf{G}_{\Omega,\Omega^c}$ was block diagonal, with blocks, $\mathbf{G}_l(\Omega)$, we see that $\sum_{l=1}^{d(\Omega)} \mathrm{rank}(\mathbf{G}_l(\Omega)) = \mathrm{rank}(\mathbf{G}_{\Omega,\Omega^c})$. If we declare an error if *any* receiver $D \in \mathcal{D}$ makes an error, we see that since we have 2^{RT} messages, from the union bound we can drive the error probability to zero if we have,

$$R < \min_{D \in \mathcal{D}} \min_{\Omega \in \Lambda_D} \mathrm{rank}(\mathbf{G}_{\Omega,\Omega^c}) \log p. \quad (21)$$

[3] Just as in the received signals, in clubbing together the transmitted signals into $\mathbf{u}_l(w)$, we put together signals transmitted at the *same* time instant together. This can be done since we have broken the network into the clusters/stages with identical path lengths.

Therefore for the layered (equal path) network with linear deterministic functions, since as seen in Section II, the cut-set is also identical to the expression in (21), we have proved the following result.

Theorem 4.1: Given a layered (equal path) linear finite-field relay network (with broadcast and multiple access), the multicast capacity C of such a relay network is given by,

$$C = \min_{D \in \mathcal{D}} \min_{\Omega \in \Lambda_D} \operatorname{rank}(\mathbf{G}_{\Omega,\Omega^c}) \log p, \tag{22}$$

V. Layered networks: general deterministic model

In this section we prove main theorems 2.1 and 2.2 for layered networks. We first generalize the encoding scheme to accommodate arbitrary deterministic functions of (1) in Section V-A. We then illustrate the ingredients of the proof using the same example as in Section III-B. Then we prove the result for layered networks in Section V-C.

A. Encoding for general deterministic model

We assume a clocked network as in Section III-A. Therefore, for such a clocked network, the deterministic model in (5) implies that

$$\mathbf{y}_j^{[t]} = g_j(\{x_i^{[t]}\}_{i \in \mathcal{N}_j}), \; t = 1, 2, \ldots, T.$$

We have a single source S with message $W \in \{1, 2, \ldots, 2^{TKR}\}$ which is encoded by the source S into a signal over KT transmission times (symbols), giving an overall transmission rate of R. We will use strong (robust) typicality as defined in [11]. The notion of joint typicality is naturally extended from Definition 5.1.

Definition 5.1: We define $\underline{x} \in T_\delta$ if

$$|\nu_{\underline{x}}(x) - p(x)| \leq \delta p(x),$$

where $\nu_{\underline{x}}(x) = \frac{1}{T}|\{t : x_t = x\}|$, is the empirical frequency.

Each relay operates over blocks of time T symbols, and uses a mapping $f_j^{[t]} : \mathcal{Y}_j^T \to \mathcal{X}_j^T$ its received symbols from the previous block of T symbols to transmit signals in the next block. In particular, block k of T received symbols is denoted by $\mathbf{y}_j^{(k)} = \{y^{[(k-1)T+1]}, \ldots, y^{[kT]}\}$ and the transmit symbols by $\mathbf{x}_j^{(k)}$. Choose some product distribution $\prod_{i \in \mathcal{V}} p(x_i)$. At the source S, map each of the indices in $W \in \{1, 2, \ldots, 2^{TKR}\}$ choose $f_S^{(k)}(W)$ onto a sequence uniformly drawn from $T_\delta(X_S)$, which is the typical set of sequences in $\mathcal{X}_S^T$. At any relay node j choose $f_j^{(k)}$ to map each typical sequence in $\mathcal{Y}_j^T$ *i.e.,* $T_\delta(Y_j)$ onto typical set of transmit sequences *i.e.,* $T_\delta(X_j)$, as

$$\mathbf{x}_j^{(k)} = f_j^{(k)}(\mathbf{y}_j^{(k-1)}), \tag{23}$$

where $f_j^{(k)}$ is chosen to map uniformly randomly each sequence in $T_\delta(Y_j)$ onto $T_\delta(X_j)$ and is done independently for each block k. Each relay does the encoding prescribed by (23). Given the knowledge of all the encoding functions $f_j^{(k)}$ at the relays and signals received over $K + |\mathcal{V}| - 2$ blocks, the decoder $D \in \mathcal{D}$, attempts to decode the message W sent by the source.

B. Proof illustration

Now, we illustrate the ideas behind the proof of Theorem 2.1 for layered networks using the same example as in Section III-B, which was done for the linear deterministic model. Since we are dealing with deterministic networks, the logic upto (10) in Section III-B remains the same. We will again illustrate the ideas using the cut $\Omega = \{S, A_1, B_1\}$. As in Section III-B, necesary condition for this set to be the distinguishability set is that $\mathbf{y}_{A_2}(w) = \mathbf{y}_{A_2}(w')$, along with $\mathbf{y}_{B_2}(w) = \mathbf{y}_{B_2}(w')$ and $\mathbf{y}_D(w) = \mathbf{y}_D(w')$. Notice that as in Section III-B, we are suppressing the block numbers associated with the received signals. It is clear that for $w = w_1$, the block numbers associated with $\mathbf{y}_{A_2}, \mathbf{y}_{B_2}, \mathbf{y}_D$ are $1, 2, 3$ respectively.

Note that since $\mathbf{y}_j \in T_\delta(Y_j)$ with high probability, we can focus only on the typical received signals. Let us first examine the probability that $\mathbf{y}_{A_2}(w) = \mathbf{y}_{A_2}(w')$. Since S can distinguish between w, w', it maps these sub-messages independently to two transmit signals $\mathbf{x}_S(w), \mathbf{x}_S(w') \in T_\delta(X_S)$, hence we can see that this probability is,

$$\mathbb{P}\left\{(\mathbf{x}_S(w'), \mathbf{y}_{A_2}(w)) \in T_\delta(X_S, Y_{A_2})\right\} = 2^{-TI(X_S; Y_{A_2})}. \tag{24}$$

Now, in order to analyze the probability that $\mathbf{y}_{B_2}(w) = \mathbf{y}_{B_2}(w')$, as seen in the linear model analysis, we see that since $\mathbf{y}_{A_2}(w) = \mathbf{y}_{A_2}(w')$, $\mathbf{x}_{A_2}(w) = \mathbf{x}_{A_2}(w')$, *i.e.,* the *same* signal is sent under both w, w'. Therefore, since naturally $(\mathbf{x}_{A_2}(w), \mathbf{y}_{B_2}(w)) \in T_\delta(X_{A_2}, Y_{B_2})$, obviously, $(\mathbf{x}_{A_2}(w'), \mathbf{y}_{B_2}(w)) \in T_\delta(X_{A_2}, Y_{B_2})$ as well. Therefore, under w', we already have $\mathbf{x}_{A_2}(w')$ to be jointly typical with the signal that is received under w. However, since A_1 can distinguish between w, w', it will map the transmit sequence $\mathbf{x}_{A_1}(w')$ to a sequence which is independent of $\mathbf{x}_{A_1}(w)$ transmitted under w. Since an error occurs when $(\mathbf{x}_{A_1}(w'), \mathbf{x}_{A_2}(w'), \mathbf{y}_{B_2}(w)) \in T_\delta(X_{A_1}, X_{A_2}, Y_{B_2})$, and since A_2 cannot distinguish between w, w', we also have $\mathbf{x}_{A_2}(w) = \mathbf{x}_{A_2}(w')$, we require that $(\mathbf{x}_{A_1}, \mathbf{x}_{A_2}, \mathbf{y}_{B_2})$

generated like $p(\mathbf{x}_{A_1})p(\mathbf{x}_{A_2}, \mathbf{y}_{B_2})$ behaves like a jointly typical sequence. Therefore, this probability is given by,

$$\mathbb{P}\left\{(\mathbf{x}_{A_1}(w'), \mathbf{x}_{A_2}(w), \mathbf{y}_{B_2}(w)) \in T_\delta(X_{A_1}, X_{A_2}Y_{B_2})\right\} \doteq 2^{-TI(X_{A_1};Y_{B_2},X_{A_2})} \overset{(a)}{=} 2^{-TI(X_{A_1};Y_{B_2}|X_{A_2})}, \quad (25)$$

where $\doteq$ indicates exponential equality (where we neglect subexponential constants), and (a) follows since we have generated the mappings f_j independently, it induces an independent distribution on X_{A_1}, X_{A_2}. Another way to see this is that the probability of (25) is given by $\frac{|T_\delta(\mathbf{X}_{A_1}|\mathbf{x}_{A_2},\mathbf{y}_{B_2})|}{|T_\delta(\mathbf{X}_{A_1})|}$, which by using properties of (robustly) typical sequences [11] yields the same expression as in (25). Note that the calculation in (25) is similar to one of the error event calculations in a multiple access channel,

Using a similar logic we can write,

$$\mathbb{P}\left\{(\mathbf{x}_{B_1}(w'), \mathbf{x}_{B_2}(w), \mathbf{y}_D(w)) \in T_\delta(X_{B_1}, X_{B_2}Y_D)\right\} \doteq 2^{-TI(X_{B_1};Y_D,X_{B_2})} \overset{(a)}{=} 2^{-TI(X_{B_1};Y_D|X_{B_2})}. \quad (26)$$

Therefore, putting (24)–(26) together as done in (14) we get

$$\mathcal{P} \leq 2^{-T\{I(X_S;Y_{A_2})+I(X_{A_1};Y_{B_2}|X_{A_2})+I(X_{B_1};Y_D|X_{B_2})\}}$$

Note that for this example, due to the Markovian structure of the network we can see that[4] $I(Y_{\Omega^c}; X_\Omega|X_{\Omega^c}) = I(X_S; Y_{A_2}) + I(X_{A_1}; Y_{B_2}|X_{A_2}) + I(X_{B_1}; Y_D|X_{B_2})$, hence as in (15) we get that,

$$P_e \leq 2^{RT}|\Lambda_D|2^{-T\min_{\Omega\in\Lambda_D} I(Y_{\Omega^c};X_\Omega|X_{\Omega^c})}, \quad (27)$$

and hence the error probability can be made as small as desired if $R < \min_{\Omega\in\Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c})$, since we are dealing with deterministic networks.

C. General deterministic model: Proof for layered networks

As in the example illustrating the proof in Section V-B, the logic of the proof in the general deterministic functions follows that of the linear model quite closely. In particular, as in Section IV we can define the bipartite network associated with a cut Ω. Instead of a transfer matrix $\mathbf{G}_{\Omega,\Omega^c}(\cdot)$ associated with the cut, we have a transfer function $\tilde{\mathbf{G}}_\Omega$. Since we are still dealing with a layered network, as in the linear model case, this transfer function breaks up into components corresponding to each of the l_D layers of the network. More precisely, we can create $d = l_D$ disjoint sub-networks of nodes corresponding to each layer of the network, with $\beta_l(\Omega)$ nodes at distance $l-1$ from S, on one side and $\gamma_l(\Omega)$ nodes at distance l from S, on the other, for $l = 1, \ldots, l_D$. Each of this MIMO clusters have a transfer function $\mathbf{G}_l(\cdot), l = 1, \ldots, l_D$ associated with them.

As in the linear model, each node $i \in \beta_l(\Omega)$ sees a signal related to $w = w_1$ in block $l_i = l-1$, and therefore waits to receive this block and then does a mapping using the general encoding function given in (23) as

$$\mathbf{x}_j^{(k)}(w) = f_j^{(k)}(\mathbf{y}_j^{(k-1)}(w)). \quad (28)$$

The received signals in the nodes $j \in \gamma_l(\Omega)$ are deterministic transformations of the transmitted signals from nodes $\mathcal{T}_l = \{u : (u, v) \in \mathcal{E}, v \in \gamma_l(\Omega)\}$. As in the linear model analysis of Section IV, the dependence is on all the transmitting signals at distance $l-1$ from the source, not just the ones in $\beta_l \subset \Omega$. Since all the receivers in γ_l are at distance l from S, they form the receivers of the MIMO layer l, and we denote this vector received signal as $\mathbf{z}_l(w)$, and this can be done for all layers $l = 1, \ldots, l_D$. Note that as in the example network of Section V-B, for all the transmitting nodes in $\mathcal{T}$ which cannot distinguish between w, w' the transmitted signal would be the same under both w and w'. Therefore, all the nodes in $\mathcal{T}_l \cap \Omega^c$ cannot distinguish between w, w' and therefore

$$\mathbf{x}_j(w) = \mathbf{x}_j(w'), \;\; j \in \mathcal{T}_l \cap \Omega^c.$$

Hence it is clear that since $(\{\mathbf{x}_j(w)\}_{j\in\mathcal{T}_l\cap\Omega^c}, \mathbf{z}_l(w)) \in T_\delta$, we have that

$$(\{\mathbf{x}_j(w')\}_{j\in\mathcal{T}_l\cap\Omega^c}, \mathbf{z}_l(w)) \in T_\delta.$$

Therefore, just as in Section V-B, we see that the probability that $\mathbf{z}_l(w) = \mathbf{z}_l(w')$, is given by,

$$\mathbb{P}\left\{\mathbf{z}_l(w) = \mathbf{z}_l(w')\right\} \doteq 2^{-TI(X_{\mathcal{T}_l\cap\Omega};Z_l,X_{\mathcal{T}_l\cap\Omega^c})}. \quad (29)$$

Since the events in each of the MIMO stages (clusters) are independent, we get that

$$\mathbb{P}\left\{\mathbf{z}_l(w) = \mathbf{z}_l(w'), l = 1, \ldots, d\right\} = \prod_{l=1}^d 2^{-TI(X_{\mathcal{T}_l\cap\Omega};Z_l,X_{\mathcal{T}_l\cap\Omega^c})} = 2^{-T\sum_{l=1}^d H(Z_l|X_{\mathcal{T}_l\cap\Omega^c})}. \quad (30)$$

Note that due to the Markovian nature of the layered network, we see that $\sum_{l=1}^d H(Z_l|X_{\mathcal{T}_l\cap\Omega^c}) = H(Y_{\Omega^c}|X_{\Omega^c})$. From this point onwards the proof closely follows the steps as in the linear model from (20) onwards. Therefore

[4] Note that though in the encoding scheme there is a dependence between $X_{A_1}, X_{A_2}, X_{B_1}, X_{B_2}$ and X_S, in the single-letter form of the mutual information, under a product distribution, $X_{A_1}, X_{A_2}, X_{B_1}, X_{B_2}, X_S$ are independent of each other. Therefore for example, Y_{B_2} is independent of X_{B_2} leading to $H(Y_{B_2}|X_{A_2}, X_{B_2}) = H(Y_{B_2}|X_{A_2})$. Using this argument for the cut-set expression $I(Y_{\Omega^c}; X_\Omega|X_{\Omega^c})$, we get the expansion.

for the layered (equal path) network with general deterministic functions we have proved the following result. Similarly in multicast scenario we declare an error if *any* receiver $D \in \mathcal{D}$ makes an error, we see that since we have 2^{RT} messages, from the union bound we can drive the error probability to zero if we have,

$$R < \max_{\prod_{i\in\mathcal{V}} p(x_i)} \min_{D\in\mathcal{D}} \min_{\Omega\in\Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c}). \tag{31}$$

Therefore we have proved the following result.

Theorem 5.2: Given a layered (equal path) general deterministic relay network (with broadcast and multiple access), we can achieve any rate R from S multicasting to all destinations $D \in \mathcal{D}$, with R satisfying:

$$R < \max_{\prod_{i\in\mathcal{V}} p(x_i)} \min_{D\in\mathcal{D}} \min_{\Omega\in\Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c}) \tag{32}$$

VI. Arbitrary networks

Given the proof for layered networks with equal path lengths, we are ready to tackle the proof of Theorem 2.1 and Theorem 2.2 for general relay networks.

The ingredients are developed below. First is that any network can be unfolded over time to create a layered deterministic network (this idea was introduced for graphs in [1] to handle cycles in a graph). The idea is to unfold the network to K stages such that i-th stage is representing what happens in the network during $(i-1)T$ to $iT-1$ symbol times. For example in figure 1(a) a network with unequal paths from S to D is shown. Figure 1(b) shows the unfolded form of this network. As we notice each node $v \in \mathcal{V}$ is appearing at stage $1 \le i \le K$ as $v[i]$. There are additional nodes: $T[i]$'s and $R[i]$'s. These nodes are just virtual transmitters and receivers that are put to buffer and synchronize the network. Since all communication links connected to these nodes ($T[i]$'s and $R[i]$'s) are modelled as wireline links without any capacity limit they would not impose any constraint on the network. One should notice that in general there must be an infinite capacity link between the same node and itself appearing at different times however, here we are omitting these links which means we limit the nodes to have a finite memory T. Now we show the following lemma,

Lemma 6.1: Assume $\mathcal{G}$ is a general deterministic network and $\mathcal{G}_{\text{unf}}^{(K)}$ is a network obtained by unfolding $\mathcal{G}$ over K time steps (as shown in figure 1). Then the following communication rate is achievable in $\mathcal{G}$:

$$R < \frac{1}{K} \max_{\prod_{i\in\mathcal{V}} p(x_i)} \min_{\Omega_{\text{unf}}\in\Lambda_D} H(Y_{\Omega_{\text{unf}}^c}|X_{\Omega_{\text{unf}}^c}) \tag{33}$$

where the minimum is taken over all cuts Ω_{unf} in $\mathcal{G}_{\text{unf}}^{(K)}$.

Proof: By unfolding $\mathcal{G}$ we get an acyclic deterministic network such that all the paths from the source to the destination have equal length. Therefore by theorem 5.2 we can achieve the rate

$$R_{\text{unf}} < \max_{\prod_{i\in\mathcal{V}} p(x_i)} \min_{\Omega_{\text{unf}}\in\Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c}) \tag{34}$$

in the time-expanded graph. Since it takes K steps to translate and achievable scheme in the time-expanded graph to an achievable scheme in the original graph, then the Lemma is proved. ■

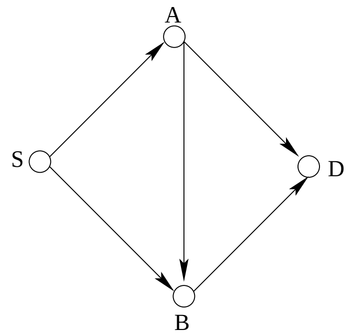
If we look at different cuts in the time-expanded graph we notice that there are two types of cuts. One type separates the nodes at different stages identically. An example of such a steady cut is drawn with solid line in figure 1 (b) which separates $\{S, A\}$ from $\{B, D\}$ at all stages. Clearly each steady cut in the time-expanded graph corresponds to a cut in the original graph and moreover its value is K times the value of the corresponding cut in the original network. However there is another type of cut which does not behave identically at different stages. An example of such a wiggling cut is drawn with dotted line in figure 1 (b). There is no correspondence between these cuts and the cuts in the original network.

Now comparing Lemma 6.1 to the main theorem 2.1 we want to prove, we notice that in this Lemma the achievable rate is found by taking the minimum of cut-values over all cuts in the time-expanded graph (steady and wiggling ones). However in theorem 2.1 we want to prove that we can achieve a rate by taking the minimum of cut-values over only the cuts in the original graph or similarly over the steady cuts in the time-expanded network. So a natural question is that in a time-expanded network does it make any difference if we take the minimum of cut-values over only steady cuts rather than all cuts ? Quite interestingly we show in the following Lemma that asymptotically as $K \to \infty$ this difference (normalized by $1/K$) vanishes.

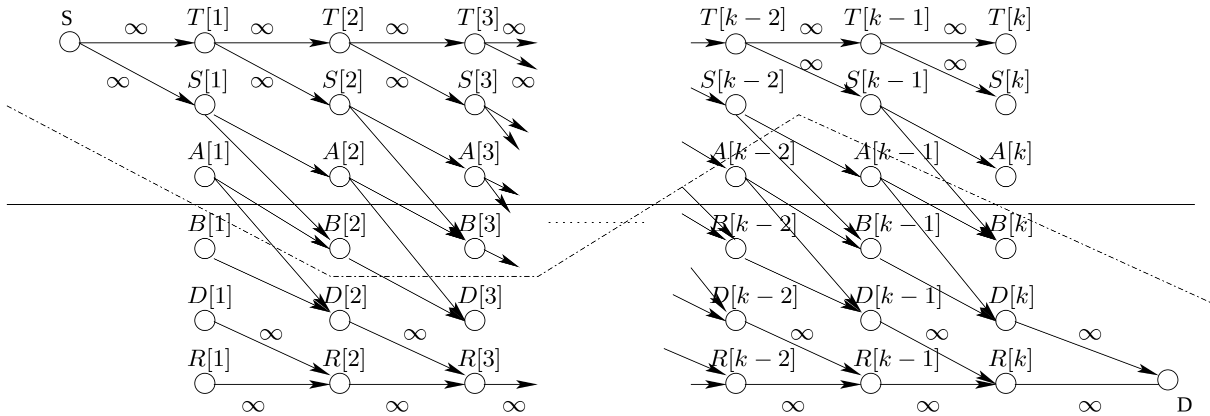
Lemma 6.2: Consider a general deterministic network, $\mathcal{G}$. Assume a product distribution on $\{x_i\}_{i\in\mathcal{V}}$, $p(\{x_i\}_{i\in\mathcal{V}}) = \prod_{i\in\mathcal{V}} p(x_i)$. Now in the time-expanded graph, $\mathcal{G}_{\text{unf}}^{(K)}$, assume that for each node $i \in \mathcal{V}$, $\{x_i[t]\}_{1\le t\le K}$ are distributed i.i.d. according to $p(x_i)$ in the original network. Also for any $1 \le t_1, t_2 \le K$ and $i \ne j$, $x_i[t_1]$ is independent of $x_j[t_2]$. Then for any cut Ω_{unf} on the unfolded graph we have,

$$(K - L + 1) \min_{\Omega\in\Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c}) \le H(Y_{\Omega_{\text{unf}}^c}|X_{\Omega_{\text{unf}}^c}) \tag{35}$$

where $L = 2^{|\mathcal{V}|-2}$.



(a) An example of general deterministic network



(b) Unfolded deterministic network. An example of steady cuts and wiggling cuts are respectively shown by solid and dotted lines.

Fig. 1. An example of a general deterministic network with un equal paths from S to D is shown in (a). The corresponding unfolded network is shown in (b).

Now since for any distribution

$$\min_{\Omega_{\text{unf}}\in\Lambda_D} H(Y_{\Omega^c_{\text{unf}}}|X_{\Omega^c_{\text{unf}}}) \leq K \min_{\Omega\in\Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c}) \quad (36)$$

we have an immediate corollary of this lemma

Corollary 6.3: Assume $\mathcal{G}$ is a general deterministic network and $\mathcal{G}^{(K)}_{\text{unf}}$ is a network obtained by unfolding $\mathcal{G}$ over K time steps then

$$\begin{aligned} &\lim_{K\to\infty} \tfrac{1}{K} \max_{\prod_{i\in\mathcal{V}} p(x_i)} \min_{\Omega_{\text{unf}}\in\Lambda_D} H(Y_{\Omega^c_{\text{unf}}}|X_{\Omega^c_{\text{unf}}}) \\ &\quad = \max_{\prod_{i\in\mathcal{V}} p(x_i)} \min_{\Omega\in\Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c}) \end{aligned} \quad (37)$$

Now by Lemma 6.1 and corollary 6.3, the proof of main theorem 2.1 is complete. So we just need to prove Lemma 6.2. First note that any cut in the unfolded graph, Ω_{unf}, partitions the nodes at each stage $1 \leq i \leq K$ to $\mathcal{U}_i$ (on the left of the cut) and $\mathcal{V}_i$ (on the right of the cut). If at one stage $S[i] \in \mathcal{V}_i$ or $D[i] \in \mathcal{U}_i$ then the cut passes through one of the infinite capacity edges (capacity Kq) and hence Lemma 6.2 is obviously proved. Therefore without loss of generality assume that $S[i] \in \mathcal{U}_i$ and $D[i] \in \mathcal{V}_i$ for all $1 \leq i \leq K$. Now since for each $i \in \mathcal{V}$, $\{x_i[t]\}_{1\leq t\leq K}$ are i.i.d distributed we can write[5]

$$H(Y_{\Omega^c_{\text{unf}}}|X_{\Omega^c_{\text{unf}}}) = \sum_{i=1}^{K-1} H(Y_{\mathcal{V}_{i+1}}|X_{\mathcal{V}_i}) \quad (38)$$

For simplification we define

$$\psi(\mathcal{V}_1,\mathcal{V}_2) \triangleq H(Y_{\mathcal{V}_2}|X_{\mathcal{V}_1}) \quad (39)$$

then we have the following lemma, whose proof is in the appendix.

Lemma 6.4: Let $\mathcal{V}_1,\ldots,\mathcal{V}_l$ be l non identical subsets of $\mathcal{V}-\{S\}$ such that $D \in \mathcal{V}_i$ for all $1 \leq i \leq l$. Also assume a product distribution on $x_i,\ i \in \mathcal{V}$. Then

$$\psi(\mathcal{V}_1,\mathcal{V}_2)+\cdots+\psi(\mathcal{V}_{l-1},\mathcal{V}_l)+\psi(\mathcal{V}_l,\mathcal{V}_1) \geq \sum_{i=1}^{l} \psi(\tilde{\mathcal{V}}_i,\tilde{\mathcal{V}}_i) \quad (40)$$

where for $k=1,\ldots,l$,

$$\tilde{\mathcal{V}}_k = \bigcup_{\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,l\}} (\mathcal{V}_{i_1}\cap\cdots\cap\mathcal{V}_{i_k}) \quad (41)$$

[5]As in Section V-B, under the product distribution the mutual information expression of the cut-set breaks into a summation.

or in another words each $\tilde{\mathcal{V}}_j$ is the union of $\binom{l}{j}$ sets such that each set is intersect of j of $\mathcal{V}_i$'s.
A special case of this Lemma was recently stated in an independent work in [14] (Lemma 2) in the context of erasure networks with only multiple access and no broadcast.

Now we are ready to prove Lemma 6.2.

Proof: (proof of Lemma 6.2) We have

$$H(Y_{\Omega^c_{\text{unf}}}|X_{\Omega^c_{\text{unf}}}) = \sum_{i=1}^{K-1} H(Y_{\mathcal{V}_{i+1}}|X_{\mathcal{V}_i}) = \sum_{i=1}^{K-1} \psi(\mathcal{V}_i, \mathcal{V}_{i+1}) \tag{42}$$

Now look at the sequence of $\mathcal{V}_i$'s. Note that there are total of $L = 2^{|\mathcal{V}|-2}$ possible subsets of $\mathcal{V}$ that contain D but not S. Assume that $\mathcal{V}_s$ is the first set that is revisited. Assume that it is revisited at step $\mathcal{V}_{s+l}$. Therefore by Lemma 6.4 we have

$$\sum_{i=1}^{l-1} \psi(\mathcal{V}_i, \mathcal{V}_{i+1}) \geq \sum_{i=1}^{l} \psi(\tilde{\mathcal{V}}_i, \tilde{\mathcal{V}}_i) \tag{43}$$

where $\tilde{\mathcal{V}}_i$'s are described in Lemma 6.4. Now note that any of those $\tilde{\mathcal{V}}_i$ contains D but not S and hence it describes a cut in the original graph, therefore $\psi(\tilde{\mathcal{V}}_i, \tilde{\mathcal{V}}_i) \geq \min_{\Omega \in \Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c})$ and hence

$$\sum_{i=1}^{l-1} \psi(\mathcal{V}_i, \mathcal{V}_{i+1}) \geq l \min_{\Omega \in \Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c}) \tag{44}$$

which means that the value of that loop is at least length of the loop times the min-cut of the original graph. Now since in any $L-1$ time frame there is at least one loop therefore except at most a path of length $L-1$ everything can be replaced with the value of the min-cut in $\sum_{i=1}^{K-1} \psi(\mathcal{V}_i, \mathcal{V}_{i+1})$. Therefore,

$$\sum_{i=1}^{K-1} \psi(\mathcal{V}_i, \mathcal{V}_{i+1}) \geq (K-L+1) \min_{\Omega \in \Lambda_D} H(Y_{\Omega^c}|X_{\Omega^c}) \tag{45}$$

■

Acknowledgements: D. Tse would like to thank Raymond Yeung for the many discussions on network coding. S. Diggavi would also like to thank Christina Fragouli for several enlightening discussions on linear network coding. The research of D. Tse and A. Avestimehr are supported by the National Science Foundation through grant CCR-01-18784 and the ITR grant:"The 3R's of Spectrum Management:Reuse, Reduce and Recycle.". The research of S. Diggavi is supported in part by the Swiss National Science Foundation NCCR-MICS center.

References

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

[2] A. S. Avestimehr, S. N. Diggavi and D. N. C. Tse, "A Deterministic Model for Wireless Relay Networks and its Capacity", *IEEE Information Theory Workshop (ITW)*, Bergen, Norway, pp 6–11, July 2007.

[3] A. S. Avestimehr, S. N. Diggavi and D. N. C. Tse, "A Deterministic Approach to Wireless Relay Networks", *Proceedings of Allerton Conference on Communication, Control, and Computing*, Illinois, September 2007.

[4] M. R. Aref, "Information Flow in Relay Networks", Ph.D. dissertation, Stanford Univ., Stanford, CA, 1980.

[5] M. Aleksic, P. Razaghi, and W. Yu, "Capacity of a Class of Modulo-Sum Relay Channels", *IEEE International Symposium on Information Theory*, Nice, France, June, 2007.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

[7] M. Effros, M. Medard, T. Ho, S. Ray, D. Karger, R. Koetter, "Linear Network Codes: A Unified Framework for Source Channel, and Network Coding", DIMACS workshop on network information theory.

[8] N J A. Harvey, R. Kleinberg, and A R. Lehman, "On the capacity of information networks", IEEE Trans. Inform. Theory, vol. 52, no. 6, pp. 2445-2464, June 2006.

[9] N. Ratnakar and G. Kramer, "The multicast capacity of deterministic relay networks with no interference", IEEE Trans. Inform. Theory, vol. 52, no. 6, pp. 2425-2432, June 2006.

[10] A. El Gamal, M. Aref, "The Capacity of the Semi-Deterministic Relay Channel", IEEE Trans. Inform. Theory, vol. 28, No. 3, p. 536, May 1982.

[11] A. Orlitsky and J. Roche, *Coding for computing*, *IEEE Transactions on Information Theory*, volume 47, number 3, pp 903–917, March 2001.

[12] P. Gupta, S. Bhadra and S. Shakkottai, "On network coding for interference networks," Procs. IEEE International Symposium on Information Theory (ISIT), Seattle, July 9-16, 2006.

[13] R. W. Yeung, A first course in information theory, Kluwer Academic/Plenum Publishers, 2002.

[14] B. Smith and S. Vishwanath, "Unicast Transmission Over Multiple Access Erasure Networks: Capacity and Duality", *IEEE Information Theory Workshop (ITW)*, Tahoe city, California, September 2007.

Appendix
Proof of Lemma 6.4

First we state a few lemmas some of whose proofs are very straightforward and hence omitted,

Lemma 1.1: The $\tilde{\mathcal{V}}_i$'s defined in Lemma 6.4 satisfy,

$$\tilde{\mathcal{V}}_l \subseteq \tilde{\mathcal{V}}_{l-1} \subseteq \cdots \subseteq \tilde{\mathcal{V}}_1 \tag{46}$$

Lemma 1.2: Let $\mathcal{V}_1, \ldots, \mathcal{V}_l$ be l non identical subsets of $\mathcal{V} - \{S\}$ such that $D \in \mathcal{V}_i$ for all $1 \leq i \leq l$. Also assume that $\tilde{\mathcal{V}}_1, \ldots, \tilde{\mathcal{V}}_l$ are as defined in lemma 6.4. Then for any $v \in \mathcal{V}$ we have

$$|\{i|v \in \mathcal{V}_i\}| = |\{j|v \in \tilde{\mathcal{V}}_j\}| \tag{47}$$

Proof: This lemma just states that for each $v \in V$ the number of times that v appears in V_i's is equal to the number of times that v appears in $\tilde{V}_i$'s. To prove it assume that v appears in V_i's is n. Then clearly

$$v \in \tilde{V}_j, \quad j = 1, \dots, n \tag{48}$$

Now for any $j > n$ any element that appears in each $\tilde{V}_j$ must appear in at least j of V_i's and since v only appears in n of V_i's therefore,

$$v \notin \tilde{V}_j, \quad j > n \tag{49}$$

therefore

$$|\{i | v \in V_i\}| = |\{j | v \in \tilde{V}_j\}| = n \tag{50}$$

■

Lemma 1.3: Let $\mathcal{V}_1, \dots, \mathcal{V}_l$ be l non identical subsets of $\mathcal{V} - \{S\}$ such that $D \in \mathcal{V}_i$ for all $1 \le i \le l$. Also assume a product distribution on $X_i,\ i \in \mathcal{V}$. Then

$$H(X_{\mathcal{V}_1}) + \cdots + H(X_{\mathcal{V}_l}) = H(X_{\tilde{\mathcal{V}}_1}) + \cdots + H(X_{\tilde{\mathcal{V}}_l}) \tag{51}$$

where $\tilde{\mathcal{V}}_i$'s are defined in Lemma 6.4 and $H(.)$ is just the binary entropy function.

Proof: For any $v \in V$ define

$$n_v = |\{i | v \in V_i\}| \tag{52}$$

and

$$\hat{n}_v = |\{j | v \in \tilde{V}_j\}| \tag{53}$$

Now since $X_i,\ i \in V$ are independent of each other we have

$$H(X_{V_1}) + \cdots + H(X_{V_l}) = \sum_{v \in V} n_v H(X_v) \tag{54}$$

and

$$H(X_{\tilde{V}_1}) + \cdots + H(X_{\tilde{V}_k}) = \sum_{v \in V} \hat{n}_v H(X_v) \tag{55}$$

By lemma 1.2 we know that $n_v = \hat{n}_v$ for all $v \in V$ hence the lemma is proved. ■

The following Lemma is just a straight forward generalization of submodularity to more than two sets (see also [8], Theorem 5 where this result is applied to the entropy function which is submodular).

Lemma 1.4: Let $\mathcal{V}_1, \dots, \mathcal{V}_k$ be a collection of sets. Assume that $\xi(\cdot)$ is a submodular function. Then,

$$\xi(\mathcal{V}_1) + \cdots + \xi(\mathcal{V}_k) \ge \xi(\tilde{\mathcal{V}}_1) + \cdots + \xi(\tilde{\mathcal{V}}_k) \tag{56}$$

where $\tilde{\mathcal{V}}_i$'s are defined in Lemma 6.4.

Now we are ready to prove Lemma 6.4. First note that

$$\psi(\mathcal{V}_1, \mathcal{V}_2) + \cdots + \psi(\mathcal{V}_{l-1}, \mathcal{V}_l) + \psi(\mathcal{V}_l, \mathcal{V}_1) =$$
$$H(Y_{\mathcal{V}_2} | X_{\mathcal{V}_1}) + \cdots + H(Y_{\mathcal{V}_l} | X_{\mathcal{V}_{l-1}}) + H(Y_{\mathcal{V}_1} | X_{\mathcal{V}_l}) =$$
$$H(Y_{\mathcal{V}_2}, X_{\mathcal{V}_1}) + \cdots + H(Y_{\mathcal{V}_l}, X_{\mathcal{V}_{l-1}}) + H(Y_{\mathcal{V}_1}, X_{\mathcal{V}_l}) - \sum_{i=1}^{l} H(X_{\mathcal{V}_i})$$

and

$$\begin{aligned} \sum_{i=1}^{l} \psi(\tilde{\mathcal{V}}_i, \tilde{\mathcal{V}}_i) &= \sum_{i=1}^{l} H(Y_{\tilde{\mathcal{V}}_i} | X_{\tilde{\mathcal{V}}_i}) && (57) \\ &= \sum_{i=1}^{l} H(Y_{\tilde{\mathcal{V}}_i}, X_{\tilde{\mathcal{V}}_i}) - \sum_{i=1}^{l} H(X_{\tilde{\mathcal{V}}_i}) && (58) \end{aligned}$$

Now define the set

$$\mathcal{W}_i = \{Y_{\mathcal{V}_i}, X_{\mathcal{V}_{i-1}}\}, \quad i = 1, \dots, l \tag{59}$$

where $\mathcal{V}_0 = \mathcal{V}_l$. Since by lemma 1.2 we have

$$\sum_{i=1}^{l} H(X_{\mathcal{V}_i}) = \sum_{i=1}^{l} H(X_{\tilde{\mathcal{V}}_i}) \tag{60}$$

we just need to prove that

$$\sum_{i=1}^{l} H(\mathcal{W}_i) \ge \sum_{i=1}^{l} H(Y_{\tilde{\mathcal{V}}_i}, X_{\tilde{\mathcal{V}}_i}) \tag{61}$$

Now by since entropy is a submodular function by Lemma 1.4 (k-way submodularity) we have,

$$\sum_{i=1}^{l} H(\mathcal{W}_i) \ge \sum_{i=1}^{l} H(\tilde{\mathcal{W}}_i) \tag{62}$$

where

$$\tilde{\mathcal{W}}_r = \bigcup_{\{i_1, \dots, i_r\} \subseteq \{1, \dots, l\}} (\mathcal{W}_{i_1} \cap \cdots \cap \mathcal{W}_{i_r}), \quad r = 1, \dots, l \tag{63}$$

Now for any r ($1 \le r \le l$) we have

$$\begin{aligned} \tilde{\mathcal{W}}_r &= \bigcup_{\{i_1, \dots, i_r\} \subseteq \{1, \dots, l\}} (\mathcal{W}_{i_1} \cap \cdots \cap \mathcal{W}_{i_r}) \\ &= \bigcup_{\{i_1, \dots, i_r\} \subseteq \{1, \dots, l\}} (\{Y_{\mathcal{V}_{i_1}}, X_{\mathcal{V}_{i_1 - 1}}\} \cap \cdots \cap \{Y_{\mathcal{V}_{i_r}} X_{\mathcal{V}_{i_r - 1}}\}) \\ &= \bigcup_{\{i_1, \dots, i_r\} \subseteq \{1, \dots, l\}} (\{Y_{\mathcal{V}_{i_1} \cap \cdots \cap \mathcal{V}_{i_r}}, X_{\mathcal{V}_{(i_1 - 1)} \cap \cdots \cap X_{\mathcal{V}_{(i_r - 1)}}}\}) \\ &= \left\{Y_{\bigcup_{\{i_1, \dots, i_r\}} (\mathcal{V}_{i_1} \cap \cdots \cap \mathcal{V}_{i_r})}, X_{\bigcup_{\{i_1, \dots, i_r\}} (\mathcal{V}_{(i_1 - 1)} \cap \cdots \cap \mathcal{V}_{(i_r - 1)})}\right\} \\ &= \{Y_{\tilde{\mathcal{V}}_r}, X_{\tilde{\mathcal{V}}_r}\} \end{aligned}$$

Therefore by equation (62) we have,

$$\begin{aligned} \sum_{i=1}^{l} H(\mathcal{W}_i) &\ge \sum_{i=1}^{l} H(\tilde{\mathcal{W}}_i) && (64) \\ &= \sum_{i=1}^{l} H(Y_{\tilde{\mathcal{V}}_i}, X_{\tilde{\mathcal{V}}_i}) && (65) \end{aligned}$$

Hence the Lemma is proved.