On Coding Capacity of Delay-constrained Network Information Flow: An Algebraic Approach

Minghua Chen
Dept. of Information Engineering
The Chinese University of Hong Kong

Ye Tian
Dept. of Information Engineering
The Chinese University of Hong Kong
Dept. of Mathematics, Nanjing University

Chih-Chun Wang
School of ECE
Purdue University

***Abstract*—Recently, Wang and Chen [1] showed that network coding (NC) can double the throughput as compared to routing in delay-constrained single-unicast communication. This is in sharp contrast to its delay-unconstrained counterpart where coding has no throughput gain. The result reveals that the landscape of delay-constrained communication is fundamentally different from the well-understood delay-unconstrained one and calls for investigation participation. In this paper, we generalize the Koetter-Medard algebraic approach [2] for delay-unconstrained network coding to the delay-constrained setting. The generalized approach allows us to systematically model deadline-induced interference, which is the unique challenge in studying network coding for delay-constrained communication. Using this algebraic approach, we characterize the coding capacity for single-source unicast and multicast, as the rank difference between an information space and a deadline-induced interference space. The results allow us to numerically compute the NC capacity for any given graph, serving as a benchmark for existing and future solutions on improving delay-constrained throughput.**

I. Introduction

Real-time communication systems that require delay guarantees have become prevalent. Typical systems of this kind include multimedia communication systems such as VoIP and video conferencing, and cyber-physical systems such as real-time surveillance and network control. As a result, delay-sensitive traffic has expressed a phenomenal growth in recent years [3]. Further, Cisco predicts that global delay-sensitive IP video will reach 104 exabytes per month in 2018, a 3-fold growth from 2013, and will account for 79% of consumer Internet traffic by then [4].

A common characteristic of these systems is that they have a strict deadline for information delivery. Information bits traversing the network need to be delivered before their deadlines, otherwise they expire and deem useless to the applications. In addition to delay constraints, real-time communication systems also require guarantees on the *timely throughput*, defined as the throughput of information bits that are delivered on time. This naturally leads to the following fundamental question:

- Given a network and an end-to-end delay constraint, how to characterize the *timely capacity*, *i.e.*, the maximum rate at which a source node can stream perishable information to its receiver nodes subject to the delay constraint?

While there have been results to the above problem under certain single-hop network settings [5]–[9], the problem remains largely open for multi-hop networks. In general, an optimal multi-hop communication scheme needs to decide the optimal routes of the information flow *in space* in order to fully utilize all the available link capacity resources, while simultaneously tracking the delay of individual packets *in time* to ensure the packets can arrive at receivers and the information can be recovered before expiration. The design problem becomes even more involved when we allow network coding [10] at intermediate nodes that intelligently mix the information content in packets before forwarding them. Such a 3-way coupling among space, time, and coding choices creates a unique challenge.

When the delay constraint is sufficiently large (*e.g.*, larger than the end-to-end delay of the longest path between the source and its receivers), the delay-unconstrained capacities are well understood. For example, for unicast, *i.e.*, from one source node to one receiver node, the delay-unconstrained capacity can be characterized by the classic min-cut/max-flow theorem, and an optimal routing solution can be obtained in polynomial time using the Ford-Fulkerson algorithm [11]. Since optimal routing already achieves the capacity, *i.e.*, the min-cut value, network coding cannot improve throughput over optimal routing when there is only one unicast flow in the network. Similarly, the research community has established a comprehensive understanding on the routing and coding capacities for delay-unconstrained broadcast and multicast, *i.e.*, from one source node to multiple receiver nodes.

The story changes completely when the delay constraint is small and is active, and our understanding of timely capacities is still nascent. Even for unicast, the timely routing and coding capacity cannot be computed by the standard graph-theoretic notion of edge cuts [12], [13], and very little is known for the cases of multicast and broadcast. Recently, Wang and Chen in [1] obtain a perhaps surprising result: there are network instances on which network coding can double the timely throughput as compared to optimal routing *even for single unicast.* This is in sharp contrast to the delay-unconstrained case where both optimal routing and coding achieve the same single-unicast capacity. Chekuri *et al.* in [14] provide an upper bound on the timely coding capacity over timely routing capacity, suggesting that coding gain in delay-constrained single-unicast may go beyond a constant value.

It is nontrivial to compute the (linear) coding capacity for delay-constrained communication. Existing results on network coding in delay-unconstrained communication do not extend directly to delay-constrained case. The new challenge lies in that in delay-constrained communication, optimal network coding needs to cancel the interference caused by future, not-yet decoded packets within the same flow. Such a new notion of interference, called *deadline-induced* interference, is strongly coupled with time (in particular deadlines) and is absent in delay-unconstrained communication.

In this paper, we provide an algebraic characterization for the timely (linear) coding capacity for single-unicast and single-multicast. In particular we generalize the well-established Koetter-Medard algebraic approach [2] for delay-unconstrained network coding to the delay-constrained setting. The generalized approach allows us to systematically model deadline-induced interference and its influence in timely coding capacity. Using this algebraic approach and leveraging an elegant technique by Guo, Cai, and Sun in [15], we characterize the coding capacity for single-source unicast and multicast, as the rank difference between an information space and a deadline-induced interference space. With proper time-expanded graph illustration similar to that in [14], our results can be thought as applying the Koetter-Medard algebraic approach to a special multi-source network coding scenario with infinite sources and a structure network of infinite size.

II. Model

Time is chopped into slots of equal length. We consider a network modeled as a directed acyclic graph $\mathcal{G} = (V, E)$, on which each edge has a capacity constraint and incurs a unit transmission delay. Links with long delay are thus modeled as a path of multiple edges.

We consider single-source delay-constrained communication scenarios where a source node s streams perishable information to a set of receiver nodes R, over the graph G. Every information bit generated at s in the *beginning* of time slot t has to be received and recovered by all receiver nodes by the *beginning* of time slot $t + D$. Here $D > 0$ is the maximum allowed end-to-end communication delay specified by applications. Since each edge incurs a unit transmission delay, with delay constraint D, any packet traverses through a path longer than D hops is deemed useless. Without loss of generality, we assume $D \leq |E|$. We follow the conventional terminology and term the communication scenario unicast if there is only one receiver node (*i.e.*, $|R| = 1$) and multicast if there is more than one receiver (*i.e.*, $|R| \leq |V| - 1$).

We consider linear network coding in delay-constrained communication on G where a vertex v can mix its incoming packets and send out coded ones. It can also delay packets before mixing them. The message a vertex sends on an outgoing edge is a linear combination of delay versions of the messages it receives. Following conventional notations [2], we name the set of all local coding coefficients and the delay factors as local encoding kernel, denoted by $\mathcal{K}$.

Let C be the minimum of min-cuts between s and receivers in R; clearly C is an upper bound for the maximum communication rate between s and receivers in R. Thus, without loss of generality, we assume that there are C outgoing edges from source s and C incoming edges into any receiver r [1]. Given the kernel $\mathcal{K}$, the network can be regarded as a linear system [2], [10]. The network transfer matrix from s to an $r \in R$, in the z-transform domain, can be expressed as

$$\phi_0^r(\mathcal{K}) + \phi_1^r(\mathcal{K})\, z^{-1} + \cdots + \phi_D^r(\mathcal{K})\, z^{-D} + \phi_{D+1}^r(\mathcal{K})\, z^{-(D+1)} + \cdots, \tag{1}$$

where every $\phi_i^r(\mathcal{K})$ is a $C \times C$ matrix that does not contain polynomials of z in its entries and is given as

$$\phi_0^r(\mathcal{K}) = 0, \text{ and } \phi_i^r(\mathcal{K}) = AF^{i-1}B_r^T,\ \forall i \geq 1.$$

Here we consider stationary linear network coding. Matrix A describes how to code the input information packets injected into the source and then place coded ones onto internal network links; the nilpotent matrix F functions as obtaining the distribution of information packets on internal network after holistic one-hop advance from the previous state; B_r is to retrieve information packets from internal network links at sink node r. To facilitate further discussion, we define

$$\Phi_r(\mathcal{K}) \triangleq \begin{bmatrix} \phi_0^r(\mathcal{K}) & \phi_1^r(\mathcal{K}) & \cdots & \phi_D^r(\mathcal{K}) \end{bmatrix}.$$

Equation (1) is insightful in that it "expands" the network transfer matrix according to the distinct delays it incurs to the messages transmitted over it. Essentially, $\phi_i^r(\mathcal{K})$ is the "network gain" corresponding to i unit delay. More specifically, for messages transmitted over the network, $\phi_i^r(\mathcal{K})$ determines the information observed by receiver r after exactly i unit delay. Thus it is conceivable that $\phi_i^r(\mathcal{K})$ and the structure revealed in (1) may be useful in studying the decoding delay of network coding, as explored in [15], [16].

In delay-unconstrained communication, given that the finite field size q is large enough, the maximum possible rate C can be achieved by optimizing the kernel $\mathcal{K}$ [2], [10] and random linear network coding can achieve the maximum rate with high probability [17].

For communication between s and r subject to delay constraint D, we leverage the insights revealed in (1) to tackle the new challenge caused by deadline-induced interference within the same information flow. In particular, we define

$$Q_D^r(\mathcal{K}) \triangleq \begin{bmatrix} \phi_0^r(\mathcal{K}) & \phi_1^r(\mathcal{K}) & \cdots & \phi_D^r(\mathcal{K}) \\ & \phi_0^r(\mathcal{K}) & \cdots & \phi_{D-1}^r(\mathcal{K}) \\ & & \ddots & \vdots \\ \mathbf{0} & & & \phi_0^r(\mathcal{K}) \end{bmatrix} \tag{2}$$

and

$$Q_{D-1}^r(\mathcal{K}) \triangleq \begin{bmatrix} \phi_0^r(\mathcal{K}) & \cdots & \phi_{D-1}^r(\mathcal{K}) \\ & \ddots & \vdots \\ \mathbf{0} & & \phi_0^r(\mathcal{K}) \end{bmatrix}. \tag{3}$$

[1] Given a graph G that does not satisfy this assumption, we can create virtual source and receiver nodes and connect them to the corresponding actual source and receiver nodes each with C parallel edges. The amended graph will satisfy the assumption and have the same timely coding capacity as G.

Let x_t be the C-dim message injected into the network and y_t^r be the C-dim message received by r, both at time t. We have

$$\begin{aligned}\left[y_0^r,\dots,y_D^r\right] &= [x_0,\dots,x_D]\, Q_D^r(\mathcal{K})\\ &= x_0\Phi_r(\mathcal{K}) + [x_1,\dots,x_D]\left[\begin{array}{cc}\mathbf{0} & Q_{D-1}^r(\mathcal{K})\end{array}\right],\end{aligned}$$

which represents what the receivers will receive from time 0 to D. Only $\left[y_0^r,\dots,y_D^r\right]$ will be used to decode x_0 since the delay requirement is D, and the contribution (or the interference) of future packets (with respect to x_0) is indeed $[x_1,\dots,x_D]\left[\begin{array}{cc}\mathbf{0} & Q_{D-1}^r(\mathcal{K})\end{array}\right]$. Thus, intuitively, $Q_D^r(\mathcal{K})$ represents the aggregate information space observed by r until delay D, and its rank represents the number of independent messages in this space. Meanwhile, $Q_{D-1}^r(\mathcal{K})$ captures the interference space spanned by future messages, and its rank represents the number of messages that got corrupted.

III. Timely (Linear) Coding Capacity

A. Unicast Case

We start from the unicast case, where s denotes the source and r denotes the only receiver.

Theorem 1. *The timely unicast (linear) coding capacity between s and r under delay constraint D can be computed by solving the following problem,*

$$\max_{\text{all possible } \mathcal{K}} \left\{rank\left(Q_D^r(\mathcal{K})\right) - rank\left(Q_{D-1}^r(\mathcal{K})\right)\right\}, \qquad (4)$$

where $Q_D^r(\mathcal{K})$ and $Q_{D-1}^r(\mathcal{K})$ are defined in (2) *and* (3), *respectively.*

Before proceeding to the proof, we make the following remarks. (i) We obtain the result by following an approach very similar to that in [15] for studying decoding delay of network codes. The key difference is that we need to address an additional challenge to guarantee the achievability of the rate computed in (4), while such challenge is absent in the problem studied in [15]. (ii) Intuitively, leveraging the understanding presented at the end of previous section on $Q_D^r(\mathcal{K})$ and $Q_{D-1}^r(\mathcal{K})$, the rank difference represents the number of "corruption-free" messages r can decode subject to delay constraint D. (iii) Theorem 1 allows us to numerically compute the timely unicast coding capacity for any given graph, serving as a benchmark for solutions on improving timely unicast throughput.

We first present a lemma to be used in the proof of Theorem 1 in the following.

Lemma 2. *For a given local coding kernel $\mathcal{K}$, the timely unicast (linear) coding capacity between s and r under delay constraint D is upper bounded by*

$$\omega \triangleq rank\left(Q_D^r(\mathcal{K})\right) - rank\left(Q_{D-1}^r(\mathcal{K})\right). \qquad (5)$$

Proof: The proof follows a set of arguments very similar to those in proving Theorem 2 in [15]. Given $\mathcal{K}$, let $\omega' \le C$ be the timely unicast linear coding capacity. Since we consider stationary linear network coding, there must exists a decoding matrix U so that ω' number of individual messages in the message vector x_0 can be decoded by time D, or equivalently,

$$Q_D^r(\mathcal{K})U = \left[\begin{array}{cc} I_{\omega'} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{array}\right].$$

This suggests that every column in $\left[\begin{array}{c} I_{\omega'}\\ \mathbf{0}\end{array}\right]$ is in the range of $Q_D^r(\mathcal{K})$; consequently, we have the following observation

$$\begin{aligned}\operatorname{rank}\left(Q_D^r(\mathcal{K})\right) &= \operatorname{rank}\left\{\left[\begin{array}{ccccc} \left[\begin{array}{c} I_{\omega'}\\ \mathbf{0}\end{array}\right] & \phi_0^r(\mathcal{K}) & \phi_1^r(\mathcal{K}) & \cdots & \phi_D^r(\mathcal{K})\\ \mathbf{0} & \mathbf{0} & \phi_0^r(\mathcal{K}) & \cdots & \phi_{D-1}^r(\mathcal{K})\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \phi_0^r(\mathcal{K})\end{array}\right]\right\}\\ &\le \operatorname{rank}\left\{\left[\begin{array}{cccc} I_{\omega'} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \phi_0^r(\mathcal{K}) & \cdots & \phi_{D-1}^r(\mathcal{K})\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \phi_0^r(\mathcal{K})\end{array}\right]\right\}\\ &= \operatorname{rank}\left(Q_{D-1}^r(\mathcal{K})\right) + \omega'.\end{aligned}$$

Thus we have $\omega' \le \operatorname{rank}\left(Q_D^r(\mathcal{K})\right) - \operatorname{rank}\left(Q_{D-1}^r(\mathcal{K})\right)$. ■

We now present the proof of Theorem 1.

Proof: By Lemma 2, we know that ω defined in (5) is an upper bound on the timely unicast capacity for a given $\mathcal{K}$. A remaining critical step is to show that this bound is indeed achievable.

We first expand $Q_{D-1}^r(\mathcal{K})$ by adding a $DC \times C$ zero matrix to its left as follows:

$$P_{D-1}^r(\mathcal{K}) \triangleq \left[\begin{array}{cc}\mathbf{0}_{DC\times C} & Q_{D-1}^r(\mathcal{K})\end{array}\right].$$

Recall that $\Phi_r(\mathcal{K}) \triangleq \left[\begin{array}{cccc}\phi_0^r(\mathcal{K}) & \phi_1^r(\mathcal{K}) & \cdots & \phi_D^r(\mathcal{K})\end{array}\right]$ is the first C rows of $Q_D^r(\mathcal{K})$. Let $m = \operatorname{rank}\left(Q_{D-1}^r(\mathcal{K})\right)$, then $\operatorname{rank}\left(P_{D-1}^r(\mathcal{K})\right) = m$ and $\operatorname{rank}\left(Q_D^r(\mathcal{K})\right)$ is $m+\omega$.

There exist m linearly independent row vectors in $P_{D-1}^r(\mathcal{K})$, denoted as $R_1,\dots,R_m$. Recall that $Q_D^r(\mathcal{K})$ and $Q_{D-1}^r(\mathcal{K})$ have the forms defined in (2) and (3), respectively. This indicates that we can find another ω row vectors from $\Phi_r(\mathcal{K})$, denoted as $S_1,\dots,S_\omega$, so that the $m+\omega$ row vectors $R_i,\dots,R_m$ and $S_1,\dots,S_\omega$ constitute a maximum linearly independent group of the row vectors in $Q_D^r(\mathcal{K})$. Take note of the row-coordinates of $\{S_i\}_{i=1}^{\omega}$ in $Q_D^r(\mathcal{K})$ as a index set $\{k_i\}_{i=1}^{\omega}$.

Let ω-dim row vector $\bar{x}_t$ denote the raw message ready to be sent from the source s in the beginning of time t. It is pre-coded into a C-dim message as follows:

$$x_t = \bar{x}_t E,$$

where $E = (e_{ij})_{\omega\times C}$ and

$$e_{ij} = \begin{cases} 1, & \text{if } j = k_i,\\ 0, & \text{otherwise.}\end{cases}$$

With such encoding, the aggregate information space (with respect to $\bar{x}_t$) observed by r until delay D is represented by

$E \circ Q_D^r(\mathcal{K})$; the interference space spanned by future messages is represented by $E \circ Q_{D-1}^r(\mathcal{K})$. Here

$$E \circ Q_D^r(\mathcal{K}) \triangleq \begin{bmatrix} E\cdot\phi_0^r(\mathcal{K}) & E\cdot\phi_1^r(\mathcal{K}) & \cdots & E\cdot\phi_D^r(\mathcal{K}) \\ & E\cdot\phi_0^r(\mathcal{K}) & \cdots & E\cdot\phi_{D-1}^r(\mathcal{K}) \\ & & \ddots & \vdots \\ 0 & & & E\cdot\phi_0^r(\mathcal{K})\end{bmatrix} \tag{6}$$

and

$$E \circ Q_{D-1}^r(\mathcal{K}) \triangleq \begin{bmatrix} E\cdot\phi_0^r(\mathcal{K}) & \cdots & E\cdot\phi_{D-1}^r(\mathcal{K}) \\ & \ddots & \vdots \\ 0 & & E\cdot\phi_0^r(\mathcal{K})\end{bmatrix}. \tag{7}$$

Next, we will show $\mathrm{rank}\left(E \circ Q_D^r(\mathcal{K})\right) - \mathrm{rank}\left(E \circ Q_{D-1}^r(\mathcal{K})\right) = \omega$. By the way we construct E, we have

$$E \cdot \Phi_r(\mathcal{K}) = \begin{bmatrix} S_1 \\ \vdots \\ S_\omega \end{bmatrix},$$

which has full row rank. Now observing that (i) $R_i, \ldots, R_m$ is a maximum linearly independent group of row vectors in $P_{D-1}^r(\mathcal{K})$ and (ii) $R_i, \ldots, R_m$ and $S_1, \ldots, S_\omega$ jointly constitute a maximum linearly independent group of the row vectors in $Q_D^r(\mathcal{K})$, we conclude that none of $S_1, \ldots, S_\omega$ can be expressed as linear combinations of row vectors in $Q_{D-1}^r(\mathcal{K})$ and thus $E \circ Q_{D-1}^r(\mathcal{K})$. As a result, the rank difference between $E \circ Q_D^r(\mathcal{K})$ and $E \circ Q_{D-1}^r(\mathcal{K})$ is also ω, which is exactly the length of $\bar{x}_t$. With this observation, applying Theorem 2 in [15], we conclude that the ω-dim vector $\bar{x}_t$ is decode-able at time $t + D$ and a timely throughput ω is achievable.

Till now, we have proved that given a local coding kernel $\mathcal{K}$, the maximum achievable timely coding throughput is given by $\mathrm{rank}\left(Q_D^r(\mathcal{K})\right) - \mathrm{rank}\left(Q_{D-1}^r(\mathcal{K})\right)$. Further maximizing over all possible coding kernel $\mathcal{K}$ gives the timely coding capacity. ■

B. Multicast Case

In multicast, there is one source but multiple receivers. Information packets for different receivers may go though different paths and were performed different coding operations. As a result, each receiver r observes a unique $Q_D^r(\mathcal{K})$. Clearly an upper bound on the timely mulicast coding capacity is the minimum of timely unicast coding capacities across all receivers $r \in R$. We show in the following theorem that it is indeed achievable.

Theorem 3. *The timely multicast (linear) coding capacity between s and R under delay constraint D can be computed by solving the following problem,*

$$\max_{\textit{all possible } \mathcal{K}} \min_{r\in R} \{rank(Q_D^r(\mathcal{K})) - rank(Q_{D-1}^r(\mathcal{K}))\}, \tag{8}$$

where $Q_D^r(\mathcal{K})$ and $Q_{D-1}^r(\mathcal{K})$ are defined in (2) *and* (3)*, respectively.*

Theorem 3 is useful in the sense that it allows us to do computer search to determine the timely multicast (linear) coding capacity for a given network. Note that to show the achieve-ability of the above capacity, we have to determine a pre-coding operation at the source so that the message can be decoded by all receivers in R. In the proof below, we construct such a common pre-coding matrix to achieve the timely multicast coding capacity in (8).

Proof: Given a local coding kernel $\mathcal{K}$, we expand $Q_{D-1}^r(\mathcal{K})$ for all $r \in R$ by supplementing a $DC \times C$ zero matrix block to its left. Let

$$\omega = \min_{r\in R} \{\mathrm{rank}(Q_D^r(\mathcal{K})) - \mathrm{rank}(Q_{D-1}^r(\mathcal{K}))\}.$$

Applying Theorem 1, since $\omega \leq \mathrm{rank}\left(Q_D^r(\mathcal{K})\right) - \mathrm{rank}\left(Q_{D-1}^r(\mathcal{K})\right)$ for all $r \in R$, we can construct find a pre-coding matrix for each receiver r, denoted as E_r, such that

$$\mathrm{rank}(E_r \circ Q_D^r(\mathcal{K})) - \mathrm{rank}(E_r \circ Q_{D-1}^r(\mathcal{K})) = \omega,$$

and also

$$\langle E_r \cdot \Phi_r \rangle \cap \langle Q_{D-1}^r(\mathcal{K}) \rangle = \{0\},$$

where $\Phi_r \triangleq \begin{bmatrix} \phi_0^r(\mathcal{K}) & \phi_1^r(\mathcal{K}) & \cdots & \phi_D^r(\mathcal{K}) \end{bmatrix}$ be the first C rows of $Q_D^r(\mathcal{K})$, $\langle\cdot\rangle$ is the linear span operator, and $E_r \circ Q_D^r(\mathcal{K})$ and $E_r \circ Q_{D-1}^r(\mathcal{K})$ are defined in a way similar to those in (6) and (7).

Each $E_r \cdot \Phi_r$ has full row rank. Thus, by adding proper rows and aggregating these new rows as a block T_r, we can construct an invertible square matrix $\begin{bmatrix} E_r \cdot \Phi_r \\ T_r \end{bmatrix}$ such that $\langle Q_{D-1}^r(\mathcal{K}) \rangle \subseteq \langle T_r \rangle$ and $\langle E_r \cdot \Phi_r \rangle \cap \langle T_r \rangle = \{0\}$. For any $r, r' \in R$, let

$$P_{r,r'} = \begin{bmatrix} S_{r,r'} & R_{r,r'} \end{bmatrix} = E_r \cdot \Phi_{r'} \begin{bmatrix} E_{r'} \cdot \Phi_{r'} \\ T_{r'} \end{bmatrix}^{-1}.$$

We have

$$\begin{aligned} E_r \cdot \Phi_{r'} &= \begin{bmatrix} S_{r,r'} & R_{r,r'} \end{bmatrix} \begin{bmatrix} E_{r'} \cdot \Phi_{r'} \\ T_{r'} \end{bmatrix} \\ &= S_{r,r'} \cdot E_{r'} \cdot \Phi_{r'} + R_{r,r'} \cdot T_{r'}, \end{aligned}$$

Note that for $r \in R$, $S_{r,r} = I_{\omega\times\omega}$ (an identity matrix) and $R_{r,r} = \mathbf{0}$.

Let $\alpha_1, \ldots, \alpha_{|R|}$ be $|R|$ finite-field variables, and we construct a common pre-coding matrix as

$$E = \sum_{r\in R} \alpha_r E_r.$$

Now we properly choose α_r in a way so that E is a pre-coding matrix that meets our requirement. Define

$$f(\alpha_1, \ldots, \alpha_{|R|}) = \prod_{r'\in R} \det\left(\sum_{r\in R} \alpha_r S_{r,r'}\right)$$

as a nonzero polynomial. Clearly the degree of α_r in the polynomial will not exceed $|R|$ from the definition. Using the same arguments on characterizing the finite-field size needed for achieving the network coding capacity under the delay-unconstrained setting in [2], we know that there exists a set of $\alpha_1, \ldots, \alpha_{|R|}$ such that $f(\alpha_1, \ldots, \alpha_{|R|}) \neq 0$, given the finite field size is strictly larger than the maximum degree of α_r in the polynomial, which is $|R|$ in this case. Suppose we have chosen $\alpha_1, \ldots, \alpha_{|R|}$ these values (so that $f(\alpha_1, \ldots, \alpha_{|R|}) \neq 0$).

We now verify that the $\omega \times C$ pre-coding matrix E constructed above allows every receiver to recover the perishable information stream of rate ω. That is to show for all $r \in R$, $\operatorname{rank}\left(E \circ Q_D^r(\mathcal{K})\right) - \operatorname{rank}\left(E \circ Q_{D-1}^r(\mathcal{K})\right) = \omega$. Equivalently, this is to show

- (i) for all $r' \in R$, $E \cdot \Phi_{r'}$ has ω linearly independent rows;
- (ii) for all $r' \in R$, $\langle E \cdot \Phi_{r'} \rangle \cap \langle E \circ Q_{D-1}^{r'} \rangle = \{0\}$.

Note that by the definition of E and $\Phi_{r'}$, we have

$$\begin{aligned} E \cdot \Phi_{r'} &= \sum_{r\in R} \alpha_r E_r \cdot \Phi_{r'} = \sum_{r\in R} \left(\alpha_r S_{r,r'} \cdot E_{r'} \cdot \Phi_{r'} + \alpha_r R_{r,r'} \cdot T_{r'}\right) \\ &= \left(\sum_{r\in R} \alpha_r S_{r,r'}\right) E_{r'} \cdot \Phi_{r'} + \left(\sum_{r\in R} \alpha_r R_{r,r'}\right) T_{r'}. \end{aligned}$$

To prove (i), we show that if there exists a vector h so that $h \cdot E \cdot \Phi_{r'} = 0$, then h must be a zero vector. Let h be a vector satisfying $h \cdot E \cdot \Phi_{r'} = 0$, then

$$h\left(\sum_{r\in R} \alpha_r S_{r,r'}\right) E_{r'} \cdot \Phi_{r'} + h\left(\sum_{r\in R} \alpha_r R_{r,r'}\right) T_{r'} = 0.$$

By the way we construct $E_{r'}$ and $T_{r'}$, we have $\langle E_{r'} \cdot \Phi_{r'} \rangle \cap \langle T_{r'} \rangle = \{0\}$. The above equation then implies

$$h\left(\sum_{r\in R} \alpha_r S_{r,r'}\right) E_{r'} \cdot \Phi_{r'} = h\left(\sum_{r\in R} \alpha_r R_{r,r'}\right) T_{r'} = 0.$$

Meanwhile, we also know that $E_{r'} \cdot \Phi_{r'}$ has full row rank and $\sum_{r\in R} \alpha_r S_{r,r'}$ is invertible and thus is also fully rank (since $f(\alpha_1, \ldots, \alpha_{|R|}) = \prod_{r'\in R} \det\left(\sum_{r\in R} \alpha_r S_{r,r'}\right) \neq 0$). Thus $h\left(\sum_{r\in R} \alpha_r S_{r,r'}\right) E_{r'} \cdot \Phi_{r'} = 0$ implies $h = 0$.

We now prove (ii) by showing that any nonzero vector in the linear space $\langle E \cdot \Phi_{r'} \rangle$ cannot be in linear space $\langle E \circ Q_{D-1}^{r'} \rangle$. Let h be a nonzero vector in $\langle E \cdot \Phi_{r'} \rangle$. We have

$$h \cdot E \cdot \Phi_{r'} = h\left(\sum_{r\in R} \alpha_r S_{r,r'}\right) E_{r'} \cdot \Phi_{r'} + h\left(\sum_{r\in R} \alpha_r R_{r,r'}\right) T_{r'}.$$

From the proof for (i) above, we know that $h\left(\sum_{r\in R} \alpha_r S_{r,r'}\right) E_{r'} \cdot \Phi_{r'} \neq 0$ and are given $\langle E_{r'} \cdot \Phi_{r'} \rangle \cap \langle T_{r'} \rangle = \{0\}$. As such, $h \cdot E \cdot \Phi_{r'}$ cannot be expressed as a linear combination of the row vectors in $T_{r'}$; consequently $h \cdot E \cdot \Phi_{r'} \notin \langle T_{r'} \rangle$. By the way we construct $T_{r'}$, we also know $\langle Q_{D-1}^{r'}(\mathcal{K}) \rangle \subseteq \langle T_{r'} \rangle$, thus $h \cdot E \cdot \Phi_{r'} \notin \langle Q_{D-1}^{r'}(\mathcal{K}) \rangle$ as well.

With (i) and (ii) proven, it is straightforward to verify that all receivers $r \in R$ have $\operatorname{rank}\left(E \circ Q_D^r(\mathcal{K})\right) - \operatorname{rank}\left(E \circ Q_{D-1}^r(\mathcal{K})\right) = \omega$, following the same arguments in the proof of Theorem 1. The proof is thus completed. ■

IV. Conclusions

There have long been interest in applying network coding in delay-sensitive applications; see some recent examples in [1], [14], [18]. Recently, wang and Chen show that network coding outperforms traditional routing by elevating the communication capacity in delay-constrained setting, even for single-unicast [1]. We continue the study initiated in [1] and give the first characterization in the literature for the timely (linear) coding capacity of single-source communication over any given network.

In the meantime, there is much room calling for future research. A remaining question in our paper is how to estimate the size of an adequate finite field. Further, even though we could numerically compute the exact timely (linear) coding capacity, the complexity of solving the corresponding combinatorial problems is high. Thus, future work could be subsequently conducted to develop easy-to-compute upper and lower bounds for timely coding capacity.

Acknolwdgement

This work was partially supported by National Science Foundation (Grant ECCS-1407603 and CCF-1422997), National Basic Research Program of China (No. 2013CB336700), and the University Grants Committee of the Hong Kong Special Administrative Region, China (Area of Excellence Grant Project No. AoE/E-02/08 and General Research Fund No. 14209115).

References

[1] C. Wang and M. Chen, "Sending perishable information: Coding improves delay-constrained throughput even for single unicast," in *Proc. IEEE ISIT*, 2014.

[2] R. Koetter and M. Médard, "An algebraic approach to network coding," *IEEE/ACM Trans. Networking*, vol. 11, no. 5, pp. 782–795, 2003.

[3] "1h 2012 global internet phenomena report," sandvine, white paper, 2012.

[4] "Cisco visual networking index: global mobile data traffic forecast update, 2013-2018," cisco, white paper, 2013.

[5] C. L. Liu and J. W. Layland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," *Journal of the ACM (JACM)*, vol. 20, no. 1, pp. 46–61, 1973.

[6] I.-H. Hou and P. Kumar, "Scheduling heterogeneous real-time traffic over fading wireless channels," in *Proc. IEEE INFOCOM*, 2010.

[7] I. Hou, P. Kumar *et al.*, "Broadcasting delay-constrained traffic over unreliable wireless links with network coding," in *Proc. ACM MobiHoc*, 2011.

[8] I.-H. Hou, V. Borkar, and P. Kumar, "A theory of qos for wireless," in *Proc. IEEE INFOCOM*, 2009.

[9] L. Deng, C.-C. Wang, M. Chen, and S. Zhao, "Timely wireless flows with arbitrary traffic patterns: Capacity region and scheduling algorithms," in *Proc. of IEEE INFOCOM*, 2016.

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

[11] L. R. Ford and D. R. Fulkerson, "Maximal flow through a network," *Canadian Journal of Mathematics*, vol. 8, no. 3, pp. 399–404, 1956.

[12] L. Ying, S. Shakkottai, A. Reddy, and S. Liu, "On combining shortest-path and back-pressure routing over multihop wireless networks," *IEEE/ACM Trans. Networking*, vol. 19, no. 3, pp. 841–854, June 2011.

[13] M. Kodialam and T. Lakshman, "On allocating capacity in networks with path length constrained routing," in *Proc. Allerton*, 2002.

[14] C. Chekuri, S. Kamath, S. Kannan, and P. Viswanath, "Delay-constrained unicast and the triangle-cast problem," in *Proc. IEEE ISIT*, 2015.

[15] W. Guo, N. Cai, and Q. T. Sun, "Time-variant decoding of convolutional network codes," *IEEE Communications Letters*, vol. 16, no. 10, pp. 1656–1659, 2012.

[16] Q. T. Sun, S. Jaggi, and S.-Y. Li, "Delay invariant convolutional network codes," in *Proc. IEEE ISIT*, 2011, pp. 2492–2496.

[17] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Trans. Information Theory*, vol. 52, no. 10, pp. 4413–4430, 2006.

[18] J. Cloud, D. Leith, and M. Médard, "A coded generalization of selective repeat arq," in *Proc. of IEEE INFOCOM*, 2015, pp. 2155–2163.