Quality-Aware Instantly Decodable Network Coding

Ye Liu and Chi Wan Sung, *Member, IEEE*

***Abstract*—In erasure broadcast channels, network coding has been demonstrated to be an efficient way to satisfy each user's demand. However, the erasure broadcast channel model does not fully characterize the information available in a "lost" packet, and therefore any retransmission schemes designed based on the erasure broadcast channel model cannot make use of that information. In this paper, we characterize the quality of erroneous packets by Signal-to-Noise Ratio (SNR) and then design a network coding retransmission scheme with the knowledge of the SNRs of the erroneous packets, so that a user can immediately decode two source packets upon reception of a useful retransmission packet. We demonstrate that our proposed scheme, namely Quality-Aware Instantly Decodable Network Coding (QAIDNC), can increase the transmission efficiency significantly compared to the existing Instantly Decodable Network Coding (IDNC) and Random Linear Network Coding (RLNC).**

***Index Terms*—Broadcast channel, Rayleigh fading, network coding, instantly decodable network coding, maximal-ratio combining.**

I. Introduction

SINCE the seminal work in [1] proved that linear network coding can achieve the capacity of multicast networks, numerous efforts have been made to demonstrate the potential benefit of network coding in various systems, such as Peer-to-Peer (P2P) networks [2], Wireless Mesh Networks (WMN) [3], Long Term Evolution Advanced (LTE-A) [4], etc. In broadcast channels, where all the users request the same set of packets from the single source, it has also been shown that linear network coding can increase system throughput significantly [5]–[7]. Joint source-channel-network coding design is also proposed to speed up the decoding process [8]. The basic idea is to make use of the packets available at different users: Since the users may receive different packets due to random loss patterns, a coded packet which linearly combines several original source packets can be more useful than an uncoded packet as the former contains more information which is useful to more users.

One way of realizing the promising gain by network coding is the use of Random Linear Network Coding (RLNC). It is shown in [9] that the capacity of multicast networks defined in [10] can be achieved with high probability, if the coefficients, used to construct network coded packets, are generated randomly from a finite field with sufficiently large size. The randomness and rateless design of RLNC ensure its great flexibility to be applied to large networks, and it has been demonstrated that RLNC can enhance throughput in peer-to-peer networks [11] and user cooperation networks [12]. However, one practical issue of RLNC is its high decoding complexity. As the decoding process involves Gaussian elimination, which is of high computational complexity, it can severely degrade end device performance [13]. Kwan *et al.* [14] proposes a way of generating network coded packets with sparse encoding vectors. Due to the sparsity of the linear system to be solved, the decoding time can be reduced. The encoding complexity, however, is much higher than that of RLNC.

The problem of decoding complexity can alternatively be solved by sacrificing throughput. By allowing a transmitting node to only perform XOR between packets [15], the decoding can be done by a receiver in linear time complexity if there is only one packet coded by the transmitter that the receiver has not received. A number of methods of generating Instantly Decodable Network Coding (IDNC) packets have been discussed in [16]–[19]. Sadeghi *et al.* [20] considers the minimization of decoding delay, so that ideally a user can decode one packet upon receiving a coded packet from the source, unlike RLNC where typically a user needs to wait much longer before being able to decode. A comparative study between IDNC and RLNC is performed in [7], showing that in general IDNC has lower throughput but may have less decoding delay against RLNC.

Many of the existing network coding solutions treat erroneous packets at the receiver side as erasures, i.e., a transmitting node regards an erroneous packet at a particular user as being completely lost. Such a treatment can be wasteful for wireless scenarios, as an erroneous packet does contain information about the original packet from the sender. There are a few exceptions, however, after Woo *et al.* proposing SOFT that allows the physical layer to pass information of each bit of a packet to higher layers in [21]. Based on the idea of SOFT, Symbol Level Network Coding (SYNC) is then proposed in [22], where the confidence values of the bits in a packet are utilized. In [23], a network coding solution using the SNRs of the packets was proposed in Demodulate-and-Forward (DmF) relay channels that utilize the information in erroneous packets at the user. Based on the Signal-to-Noise Ratios (SNRs) of the packets in error, the relay retransmits network coded packets and a significant improvement on retransmission efficiency is observed.

In this paper, we extend the idea in [23] to wireless broadcast channels. We propose Quality-Aware Network Coding (QANC), which defines how network coding can be performed

Manuscript received June 10, 2013; revised October 6 and November 28, 2013; accepted November 28, 2013. The associate editor coordinating the review of this paper and approving it for publication was D. Niyato.

The authors are with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: yeliu27-c@my.cityu.edu.hk, albert.sung@cityu.edu.hk).

This work was supported in part by a grant from the University Grants Committee of the Hong Kong Special Administrative Region, China, under Project AoE/E-02/08, and in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project CityU 121713.

Digital Object Identifier 10.1109/TWC.2014.012314.131046

given the SNRs of the packets when Phase-Shift Keying (PSK) modulations are applied. The advantage of QANC against conventional network coding is that a user may decode two of its requesting packets upon receiving a coded packet, whereas in conventional network coding schemes a user can only decode at most one requested packet. We then propose an algorithm, namely Quality-Aware Instantly Decodable Network Coding (QAIDNC), which aims at serving the users' requests as quickly as possible using QANC. We are able to show that:

1. After careful analysis, we discover that SYNC is not suitable for the broadcasting task considered in this paper. We demonstrate that even when there is only 1 user to serve, SYNC has to send strictly more than Hybrid Automatic Repeat-Request using Chase Combining (CC-HARQ) [24], while QAIDNC can have better performance than CC-HARQ.
2. Our proposed QAIDNC scheme outperforms existing Instantly decodable Network Coding (IDNC) [16] schemes significantly while retaining the advantages of IDNC, namely low encoding complexity, low decoding complexity, and low decoding delay.
3. We show that in a wide range of number of users, QAIDNC outperforms RLNC significantly. This makes QAIDNC a desirable choice under those scenarios as RLNC has large decoding complexity and decoding delay.
4. Although QAIDNC depends on the feedback of the SNRs from the users, we show through simulations that it is quite robust against feedback loss. The performance degradation due to quantization of the feedback SNRs is also not significant.

The rest of the paper is organized as follows: Section II introduces the system model this paper concerns and some useful definitions. Section III derives the idea of QANC. Section IV gives the general problem formulation using QANC with some motivating examples. Section V describes the detailed design of QAIDNC with encoding and decoding complexity analysis. Simulation results and the conclusion are given in Section VII and Section VIII, respectively.

II. System Model and Some Definitions

A. System Model

Consider the situation shown in Fig. 1 where a source node wishes to deliver N *source packets* to K users through independent Rayleigh fading channels, and the objective is to use the minimum number of transmissions before all the users receive all the source packets correctly. Let $\mathcal{N} \triangleq \{1, 2, \ldots, N\}$ and $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$ be the index set of the N source packets and K users, respectively. Assume the transmissions performed by the source node are equally time slotted. A source packet $\mathbf{P}_n$ is a ζ-dimensional vector of symbols in $GF(q)$, where $q = 2^\eta \geq K,\ \eta \in \mathbb{Z}^+$. Since each symbol in $GF(2^\eta)$ can be represented by an η-dimensional vector over $GF(2)$, $\mathbf{P}_n$ can equivalently be represented by a vector of $\zeta\eta$ bits, denoted by $\mathbf{P}_n^{bit}$.

Assume the transmissions experience independent flat Rayleigh fading. Denote the channel coefficient of the channel link from the source to the kth user at time slot n as $h_{k,n}$, where the real and imaginary parts of $h_{k,n}$ are i.i.d. zero mean Gaussian random variables with variance 0.5 and remain constant during that time slot. At the user side, a received packet is corrupted by Additive White-Gaussian Noise (AWGN) with zero mean and σ^2 variance. Assume the source's transmission power is 1, the Signal-to-Noise Ratio (SNR) of the received packet at user k during time slot n, denoted as $\Gamma_{k,n}$, is $\frac{|h_{k,n}|^2}{\sigma^2}$. It is assumed that the users know the Channel-State Information (CSI) but the source does not.

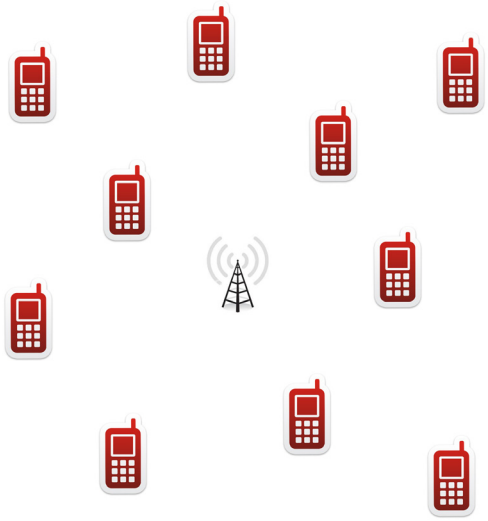


Fig. 1. The system model considered in this paper. A source in the middle is trying to broadcast data packets to all the users.

A packet is said to be received correctly if the Symbol-Error Rate (SER) of the received packet is equal to or less than a threshold denoted as ε_{th}. Given fixed modulation and coding schemes, the SER requirement can be translated into an SNR requirement. We assume that the packet is received correctly if its SNR at a user exceeds a threshold value, denoted by T.

The task is divided into two phases, namely initial phase and retransmission phase. In the initial phase, the source broadcasts the N source packets to all the users. Starting with the first packet, the source broadcasts the packet to the K users and then waits for user feedback. The feedback channels are assumed to be error-free, and the users give feedback on the SNRs of the received packets to the source. The process is repeated for each source packet. After the initial phase, the source will have the SNR Feedback Matrix (SFM), denoted as $\mathbf{\Phi}$, where $\Gamma_{k,n}$ denotes the element in the kth row and nth column, telling the SNR of $\mathbf{P}_n$ at user k.

If all the elements in $\mathbf{\Phi}$ are larger than or equal to T, then the delivery of the N source packets to the K users is done. Otherwise, the retransmission phase will be carried out until every user's demand is satisfied.

B. Definitions

In this section we give some definitions which facilitate later discussions.

Encoding vector: Denoted as $\mathbf{v}$, it is a vector of length N used to describe a linearly network coded packet. Each

element in $\mathbf{v}$ is taken from $GF(q)$, and the packet to be transmitted is calculated as $\mathbf{v}(1)\mathbf{P}_1+\mathbf{v}(2)\mathbf{P}_2+\cdots+\mathbf{v}(N)\mathbf{P}_N$, where $\mathbf{v}(i)$ denotes the ith element in $\mathbf{v}$.

Innovative: A linearly coded packet is innovative to user k, if its encoding vector is linearly independent of all the encoding vectors user k has received so far.

Worthless: A packet is said to be worthless to user k, if the packet is formed by a set of source packets that have been already decoded by user k.

Instantly decodable: A packet is instantly decodable to user k, if user k can decode 1 source packet immediately upon the reception of the linear combination.

2-instantly decodable: A packet is 2-instantly decodable to user k, if user k can decode 2 source packets immediately upon the reception of the packet.

Note that when transmitting a linearly network coded packet, the corresponding encoding vector must be included in the packet header. The overhead created by adding the encoding vectors, however, is small compared to the length of a packet in a practical system [25].

III. Quality-Aware Network Coding

Consider the case where $K = 1$, $N = 2$, and $\mathbf{\Phi} = [\Gamma_{1,1}\ \Gamma_{1,2}]$, where $\Gamma_{1,1}, \Gamma_{1,2} < T$. It is shown in [26] that if $\Gamma_{1,1} + \Gamma_{1,2} \geq T$, then the user can decode both $\mathbf{P}_1$ and $\mathbf{P}_2$ upon receiving $\mathbf{P}_1^{bit} \oplus \mathbf{P}_2^{bit}$, given that Binary Phase-Shift Keying (BPSK) is used as the modulation scheme. Here we first describe the idea in [26] for completeness and then extend it to M-ary Phase-Shift Keying (PSK) schemes, assuming $M = 2^\alpha$, $\alpha \in \mathbb{Z}^+$, and α divides $\zeta\eta$.

A. BPSK Case

In the initial phase, the source modulates $\mathbf{P}_1$ and $\mathbf{P}_2$ into $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. The source then transmits $\mathbf{x}_1$ and $\mathbf{x}_2$ in two consecutive time slots. Let the user receive $\mathbf{y}_1 = h_1\mathbf{x}_1 + \mathbf{z}_1$ and $\mathbf{y}_2 = h_2\mathbf{x}_2 + \mathbf{z}_2$, where $\mathbf{z}_1$ and $\mathbf{z}_2$ are AWGN terms.

Let the source retransmit $\mathbf{P}_1^{bit} \oplus \mathbf{P}_2^{bit}$, where $\oplus$ is the bit-wise XOR operation. Let bit 0 be modulated to symbol "-1" and bit 1 be modulated to symbol "1". It can be seen that $\mathbf{P}_1^{bit} \oplus \mathbf{P}_2^{bit}$ will be modulated to $\mathbf{x}_1 \star \mathbf{x}_2$, where $\star$ denotes symbol-wise multiplication. Note that for a given sequence of BPSK symbols $\mathbf{x}$, $\mathbf{x} \star \mathbf{x}$ is equal to the sequence of symbols all equal to symbol "1". Assume $\mathbf{x}_1 \star \mathbf{x}_2$ is correctly received by the user. By the following process:

$$\begin{aligned} \mathbf{y}_2 \star \mathbf{x}_1 \star \mathbf{x}_2 &= (h_2\mathbf{x}_2 + \mathbf{z}_2) \star (\mathbf{x}_1 \star \mathbf{x}_2) \\ &= h_2\mathbf{x}_2 \star \mathbf{x}_1 \star \mathbf{x}_2 + \mathbf{x}_1 \star \mathbf{x}_2 \star \mathbf{z}_2 \\ &= h_2\mathbf{x}_1 + \mathbf{x}_1 \star \mathbf{x}_2 \star \mathbf{z}_2, \end{aligned} \quad (1)$$

a new noisy observation of $\mathbf{P}_1$ is obtained. As the power of the AWGN term in (1) remains unchanged, this new noisy observation of $\mathbf{P}_1$ has SNR $\Gamma_{1,2}$. Using Maximal Ratio Combining (MRC) [27] to combine $\mathbf{y}_1$ with $\mathbf{y}_2 \star \mathbf{x}_1 \star \mathbf{x}_2$, we have a noisy observation of $\mathbf{P}_1$ with SNR $\Gamma_{1,1} + \Gamma_{1,2} \geq T$, which means that user 1 can correctly decode $\mathbf{P}_1$. With $\mathbf{P}_1$ and $\mathbf{P}_1^{bit} \oplus \mathbf{P}_2^{bit}$, the user can then derive $\mathbf{P}_2$.

B. M-ary PSK Case

Let the constellation points of M-ary PSK be $U_M = \{e^{j2\pi\beta/M} :\ \beta = 0, 1, \ldots, M-1\}$, where $j^2 = -1$. Note that U_M is an Abelian group under complex multiplication [28], with the point 1 (i.e., when $\beta = 0$) as its identity element. We denote the inverse of $e^{j2\pi\beta/M}$ in U_M as $(e^{j2\pi\beta/M})^{-1}$. In other words, we have $e^{j2\pi\beta/M}(e^{j2\pi\beta/M})^{-1} = 1$.

Let $\alpha = \log_2 M$. Divide $\mathbf{P}_1^{bit}$ and $\mathbf{P}_2^{bit}$ into $\zeta\eta/\alpha$ consecutive bit sequences, namely $[\mathbf{p}_{1,1}\ \mathbf{p}_{1,2} \cdots \mathbf{p}_{1,\zeta\eta/\alpha}]$ and $[\mathbf{p}_{2,1}\ \mathbf{p}_{2,2} \cdots \mathbf{p}_{2,\zeta\eta/\alpha}]$, respectively. Let f be the modulation function that maps an α-bit sequence to a symbol in U_M using gray mapping. Note that f is a bijective mapping, and we denote its inverse function as f^{Inv}.

In the initial phase, $\mathbf{P}_1^{bit}$ and $\mathbf{P}_2^{bit}$ are modulated to the following two symbol sequences $\mathbf{x}_1 = [x_{1,1}\ x_{1,2} \cdots x_{1,\zeta\eta/\alpha}]$ and $\mathbf{x}_2 = [x_{2,1}\ x_{2,2} \cdots x_{2,\zeta\eta/\alpha}]$, where

$$x_{i,r} = f(\mathbf{p}_{i,r}) = e^{j2\pi\beta_{i,r}/M},$$

for $i = 1, 2$ and $r = 1, 2, \ldots, \zeta\eta/\alpha$. For notational simplicity, we define $\mathbf{f}(\mathbf{P}_i) \triangleq [f(\mathbf{p}_{i,1})\ f(\mathbf{p}_{i,2}) \cdots f(\mathbf{p}_{i,\zeta\eta/\alpha})] = \mathbf{x}_i$. Let $\mathbf{f}^{Inv}(\mathbf{x}_i) = \mathbf{P}_i$ be the inverse function of $\mathbf{f}(\mathbf{P}_i)$, which can be obtained by applying f^{Inv} to each of the components of $\mathbf{x}_i$. Furthermore, we define

$$(\mathbf{x}_i)^{-1} \triangleq [(\mathbf{x}_{i,1})^{-1}\ (\mathbf{x}_{i,2})^{-1}\ \cdots\ (\mathbf{x}_{i,\zeta\eta/\alpha})^{-1}].$$

Note that $(\mathbf{x}_i)^{-1}$ should not be confused with $\mathbf{f}^{Inv}(\mathbf{x}_i)$: the former one is a symbol vector while the latter is a packet.

After the initial phase, the user receives $\mathbf{y}_n = [y_{i,1}\ y_{i,2} \cdots y_{i,\zeta\eta/\alpha}]$ in slot n, $n = 1, 2$, where

$$y_{n,i} = h_n e^{j2\pi\beta_{n,i}/M} + z_{n,i},$$

for $i = 1, 2, \ldots, \zeta\eta/\alpha$. Assume again $\Gamma_{1,1}, \Gamma_{1,2} < T$ and $\Gamma_{1,1} + \Gamma_{1,2} \geq T$.

Without loss of generality, we consider the first symbol of $\mathbf{y}_1$ and $\mathbf{y}_2$, i.e., $y_{1,1}$ and $y_{2,1}$. In the retransmission phase, let the source transmit the symbol $(f(\mathbf{p}_{1,1}))^{-1}f(\mathbf{p}_{2,1})$ and the user receive it correctly. Multiplying it with $y_{1,1}$, the user obtain

$$\begin{aligned} \tilde{y}_{2,1} &= (f(\mathbf{p}_{1,1}))^{-1}f(\mathbf{p}_{2,1})y_{1,1} \\ &= (e^{j2\pi\beta_{1,1}/M})^{-1}e^{j2\pi\beta_{2,1}/M}\left[h_1 e^{j2\pi\beta_{1,1}/M} + z_{1,1}\right] \\ &= h_1 e^{j2\pi\beta_{2,1}/M} + (e^{j2\pi\beta_{1,1}/M})^{-1}e^{j2\pi\beta_{2,1}/M}z_{1,1}. \end{aligned} \quad (2)$$

Hence, we get another noisy observation of $\mathbf{p}_{2,1}$ with SNR $\Gamma_{1,1}$. Combining $\tilde{y}_{2,1}$ with $y_{2,1}$ using MRC, we can get a noisy observation of $\mathbf{p}_{2,1}$ with SNR $\Gamma_{1,1} + \Gamma_{1,2} \geq T$, so that $\mathbf{p}_{2,1}$ can be correctly decoded. With $\mathbf{p}_{2,1}$ and $(f(\mathbf{p}_{1,1}))^{-1}f(\mathbf{p}_{2,1})$, we can derive $\mathbf{p}_{1,1}$ by the following steps:

1. Multiply $(f(\mathbf{p}_{1,1}))^{-1}f(\mathbf{p}_{2,1})$ with $(f(\mathbf{p}_{2,1}))^{-1}$ and get $(f(\mathbf{p}_{1,1}))^{-1}$.
2. Find the inverse of $(f(\mathbf{p}_{1,1}))^{-1}$ in U_M, which is $f(\mathbf{p}_{1,1})$.
3. Find $\mathbf{p}_{1,1} = f^{Inv}(f(\mathbf{p}_{1,1}))$.

From the above derivations, we can see that the source can send the $\zeta\eta/\alpha$-dimensional vector $(\mathbf{f}(\mathbf{P}_1))^{-1} \star \mathbf{f}(\mathbf{P}_2)$. Upon its correct reception, the user can repeatedly carry out the aforementioned procedure for each of its $\zeta\eta/\alpha$ components, so that $\mathbf{P}_1$ and $\mathbf{P}_2$ can be decoded.

In what follows, we discuss the application of the above idea to a more general setting.

Definition 1. *Let $\mathcal{A},\ \mathcal{B} \subset \mathcal{N}$, and $\mathcal{A} \cap \mathcal{B} = \emptyset$. The modulation inverse of packets in $\mathcal{A}$ coded with packets in $\mathcal{B}$, denoted as $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$, is given as*

$$\mathbf{P}_{\mathcal{A}\odot\mathcal{B}} \triangleq \mathbf{f}^{Inv}[\bigstar_{a\in\mathcal{A}}(\mathbf{f}(\mathbf{P}_a))^{-1} \star \bigstar_{b\in\mathcal{B}}\mathbf{f}(\mathbf{P}_b)]$$

where

$$\bigstar_{a\in\mathcal{A}}\mathbf{x}_a \triangleq \mathbf{x}_1 \star \mathbf{x}_2 \star \cdots \star \mathbf{x}_n,\ \mathcal{A} = \{1, 2, ..., n\}.$$

Note that the above operation on $\mathcal{A}$ and $\mathcal{B}$ is not commutative. The following result shows a condition under which a source can transmit the packet $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ to help a user to decode two packets:

Proposition 1. *Let $\mathcal{A},\ \mathcal{B} \subset \mathcal{N}$, where $\mathcal{A} \cap \mathcal{B} = \emptyset$. Also let $a \in \mathcal{A}$, $b \in \mathcal{B}$ where $\Gamma_{k,a} < T$, $\Gamma_{k,b} < T$, and $\Gamma_{k,i} \geq T$, $\forall i \in \mathcal{A} \cup \mathcal{B} \setminus \{a, b\}$. If $\Gamma_{k,a} + \Gamma_{k,b} \geq T$, then user k can decode $\mathbf{P}_a$ and $\mathbf{P}_b$ upon correctly receiving $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$.*

Proof: Since user k has decoded all source packets with indices in $\mathcal{A} \cup \mathcal{B} \setminus \{a, b\}$, she can find out

$$\mathbf{x}' = \bigstar_{i\in\mathcal{A}\setminus\{a\}}(\mathbf{f}(\mathbf{P}_i))^{-1} \star \bigstar_{i\in\mathcal{B}\setminus\{b\}}\mathbf{f}(\mathbf{P}_i).$$

Observe that for any packet $\mathbf{P}$, the operation $\mathbf{f}(\mathbf{P}) \star (\mathbf{f}(\mathbf{P}))^{-1}$ gives a vector of symbols all equal to the identity in U_M, we have:

$$\mathbf{x}' \star \mathbf{f}(\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}) = (\mathbf{f}(\mathbf{P}_a))^{-1} \star \mathbf{f}(\mathbf{P}_b).$$

According to what we have just discussed, $\mathbf{P}_a$ and $\mathbf{P}_b$ can be decoded. ∎

Note that $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ can be treated as a non-linear combination of packets with indices in $\mathcal{A}$ and packets with indices in $\mathcal{B}$, when the underlying modulation scheme is M-ary PSK, for $M = 2^\alpha > 2,\ \alpha \in \mathbb{Z}^+$. For BPSK, $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ reduces to $\mathbf{P}^{bit}_{a_1} \oplus \mathbf{P}^{bit}_{a_2} \oplus \cdots \mathbf{P}^{bit}_{a_m} \oplus \mathbf{P}^{bit}_{b_1} \oplus \mathbf{P}^{bit}_{b_2} \cdots \oplus \mathbf{P}^{bit}_{b_n}$, where $\mathcal{A} = \{a_1, a_2, ..., a_m\}$, $\mathcal{B} = \{b_1, b_2, ..., b_n\}$. To inform a user that the incoming packet is $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ when a modulation scheme higher than BPSK is used, the packet header should indicate whether a source packet belongs to $\mathcal{A}$, to $\mathcal{B}$, or to none of them. Hence, the header length would be about $N \log_2 3$ bits, which is similar to the length of an encoding vector, $N \log_2 q$ bits.

IV. Problem Formulation and Lower Bound

A. Problem Formulation

Inspired by the promising gain offered by quality-aware network coding, we aim at increasing the efficiency of retransmissions by attempting on reducing the number of retransmissions with the knowledge of $\mathbf{\Phi}$, assuming that the retransmissions are error-free. Let

$$\mathcal{Q}^{\mathrm{LNC}} \triangleq \{\sum_{i\in\mathcal{N}} \beta_i \mathbf{P}_i :\ \beta_i \in GF(q)\},$$

and

$$\mathcal{Q}^{\mathrm{QANC}} \triangleq \{\mathbf{P}_{\mathcal{A}\odot\mathcal{B}} :\ \mathcal{A} \subset \mathcal{N},\ \mathcal{B} \subset \mathcal{N},\ \mathcal{A} \cap \mathcal{B} = \emptyset\},$$

where LNC stands for linear network coding. The first set corresponds to the use of linear network coding, where retransmitted packets are obtained by linear combinations of source packets over GF(q). The second set corresponds to the use of quality-aware network coding, where retransmitted packets are obtained by the coding operation in Definition 1. How to choose the retransmitted packets can be formulated as an optimization problem as follows:

Problem 1. *Given a $K \times N$ real matrix $\mathbf{\Phi}$ and a real number T, minimize the number of retransmissions by sending packets in $\mathcal{Q}^{\mathrm{LNC}} \cup \mathcal{Q}^{\mathrm{QANC}}$, such that each user can decode $\mathbf{P}_1,\ \mathbf{P}_2,\ ...,\ \mathbf{P}_N$ in the following two ways:*

[Dec-1.] With the knowledge of some retransmitted packets in $\mathcal{Q}^{\mathrm{LNC}}$ and source packets already received, decode source packets by solving a system of linear equations that describes the linear combinations of those packets.

[Dec-2.] With the knowledge of $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ and source packets already received by user k, decode $\mathbf{P}_a$ and $\mathbf{P}_b$, where $a \in \mathcal{A}$, $b \in \mathcal{B}$, provided that $\Gamma_{k,a} + \Gamma_{k,b} \geq T$ and $\Gamma_{k,i} \geq T$ for all $i \in \mathcal{A} \cup \mathcal{B} \setminus \{a, b\}$.

When BPSK modulation is applied, $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ reduces to a linear combination of packets with indices in $\mathcal{A}$ and $\mathcal{B}$. In that case, the source will always send linear combinations of a subset of the N source packets. However, when higher modulation schemes are used, the source may send non-linear combinations in order to utilize the quality information for higher retransmission efficiency.

B. Lower Bound

Finding an optimum solution to the above problem is involved. Instead, we will now give a lower bound on the minimum number of retransmissions required in the following theorem. Let $W(\mathbf{\Phi})$ be the optimum solution to Problem 1, and $W^{\mathrm{LNC}}(\mathbf{\Phi})$ be the optimal solution using linear network coding only, i.e. only packets in $\mathcal{Q}^{\mathrm{LNC}}$ are used for retransmissions. The k-th row of $\mathbf{\Phi}$ is denoted by $\mathbf{\Phi}_k$.

Theorem 1. *The following inequalities hold for any $\mathbf{\Phi}$:*

$$\max\left\{\max_{k\in\mathcal{K}} W(\mathbf{\Phi}_k), W^{\mathrm{LNC}}(\mathbf{\Phi})/2\right\} \leq W(\mathbf{\Phi}) \leq W^{\mathrm{LNC}}(\mathbf{\Phi}).$$

Proof: The second inequality is obvious, since forbidding the use of $\mathcal{Q}^{\mathrm{QANC}}$ can only increase the number of retransmissions required. The first inequality consists of two parts. We first prove that

$$\max_{k\in\mathcal{K}} W(\mathbf{\Phi}_k) \leq W(\mathbf{\Phi}).$$

This result can be interpreted as a single-user bound. Its validity can be seen as $W(\mathbf{\Phi})$ can satisfy any user in $\mathcal{K}$, while $W(\mathbf{\Phi}_k)$ does not necessarily satisfy a user in $\mathcal{K} \setminus \{k\}$. The second part follows from the fact that using QANC can lead to at most 50% of reductions on minimum number of retransmissions, as a packet from $\mathcal{Q}^{\mathrm{QANC}}$ can help a user decode at most two source packets. ∎

Note that $W^{\mathrm{LNC}}(\mathbf{\Phi})$ is easy to find if $q \geq K$. For that case, it is optimal to always transmit innovative vectors. $W^{\mathrm{LNC}}(\mathbf{\Phi})$ can then be obtained by counting the number of entries less than T in each row, and then find the maximum. The

complexity is $O(KN)$. The other lower bound can also be found efficiently. It has been proven in [23, Theorem 1] that $W(\mathbf{\Phi}_k)$ can be found with complexity $O(N \log N)$. To find $\max_{k\in\mathcal{K}} W(\mathbf{\Phi}_k)$, we can first find $W(\mathbf{\Phi}_k)$ for each $k \in \mathcal{K}$ and then find the maximum. The total complexity is therefore $O(KN \log N)$.

C. An Illustrative Example

We use the following example for illustration of the potential gain achievable by applying QANC.

Example 1. *Suppose* $K = 2,\ N = 4,\ T = 1$, *and*

$$\mathbf{\Phi} = \begin{bmatrix} 0.5 & 0.5 & 1 & 1 \\ 1 & 1 & 0.7 & 0.3 \end{bmatrix}.$$

The source can send $\mathbf{P}_{\{1,3\}\odot\{2,4\}}$. *After receiving this packet, user 1 can calculate* $\mathbf{P}_{\{1\}\odot\{2\}}$ *based on her knowledge of* $\mathbf{P}_3$ *and* $\mathbf{P}_4$. *After that,* $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *can be recovered. Similarly, user 2 can obtain* $\mathbf{P}_3$ *and* $\mathbf{P}_4$. *In this example, one retransmission is enough. If the users treat erroneous packets as erasures, then the source needs to retransmit at least two packets (e.g.* $\mathbf{P}_1+\mathbf{P}_3$ *and* $\mathbf{P}_2+\mathbf{P}_4$*). The use of quality-aware network coding yields a reduction of 50%.*

V. Quality-Aware Instantly Decodable Network Coding Retransmissions

In this section, we propose Quality-Aware Instantly Decodable Network Coding (QAIDNC). A retransmitted packet in QAIDNC is required to satisfy one of the following three conditions:

(C.1) 2-instantly decodable to a user, or
(C.2) Instantly decodable to a user, or
(C.3) Worthless to a user.

Also, a user will try to utilize an erroneous received packet which is not worthless to her to speed up the decoding process:

M-1. If the retransmitted packet is a source packet, then the user will use MRC to combine all copies of the respective source packet.
M-2. If the retransmitted packet is a coded packet, then:
  1) If the user has seen the same packet before, she will apply MRC to combine all copies of the coded packet. If the coded packet can be decoded (and therefore the involved source packets), the user will clear the coded packet from its memory. Otherwise, the user will continue to store the coded packet.
  2) If the user has not seen the coded packet before and the code packet cannot be correctly decoded, she will store the coded packet.
M-3. When a user decodes some source packets, she checks whether any of the coded packets previously stored can be transformed into source packets that are yet to be decoded. If so, she will apply MRC to all the copies of the source packet.

The next example illustrates how a user may utilize its stored coded packets as described in M-2 and M-3.

Example 2. *Suppose* $K = 2,\ N = 3,\ T = 1$, *and*

$$\mathbf{\Phi} = \begin{bmatrix} 0.5 & 0.5 & 0.4 \\ 0.3 & 0.8 & 0.6 \end{bmatrix}.$$

Let user 1 have stored $\mathbf{P}_{\{1\}\odot\{2\}}$ *and* $\mathbf{P}_{\{2\}\odot\{3\}}$ *with SNR 0.5 and 0.7, respectively. Let user 2 have stored* $\mathbf{P}_{\{1\}\odot\{2\}}$ *and* $\mathbf{P}_{\{3\}\odot\{2\}}$ *with SNR 0.6 and 0.5, respectively. Now assume the source retransmits* $\mathbf{P}_{\{1\}\odot\{2\}}$, *and both users find out that they receive a new copy of* $\mathbf{P}_{\{1\}\odot\{2\}}$ *with SNR 0.9. By applying MRC to different copies of* $\mathbf{P}_{\{1\}\odot\{2\}}$, *both users will have* $\mathbf{P}_{\{1\}\odot\{2\}}$ *with SNRs exceeding* T, *and therefore they can now decode* $\mathbf{P}_1$ *and* $\mathbf{P}_2$.

At this point, user 1 has $\mathbf{P}_2$, *a noisy copy of* $\mathbf{P}_3$ *with SNR 0.4, and a noisy copy of* $\mathbf{P}_{\{2\}\odot\{3\}}$ *with SNR 0.7. She can now derive a new copy of* $\mathbf{P}_3$ *with SNR 0.7, as demonstrated in* (2). *By applying MRC to the two copies of* $\mathbf{P}_3$, *she gets a new copy of* $\mathbf{P}_3$ *with SNR 1.1, exceeding* T. *Therefore, user 1 now decodes all the 3 packets.*

The situation of user 2 depends on the modulation scheme used. If BPSK modulation is applied, then $\mathbf{P}_{\{3\}\odot\{2\}}$ *is equivalent to* $\mathbf{P}_2 \oplus \mathbf{P}_3$, *and she can derive a new copy of* $\mathbf{P}_3$ *with SNR 0.5 as shown in* (1). *However, if a higher PSK modulation scheme is used, user 2 will not be able to derive a new copy of* $\mathbf{P}_3$, *as the noisy copy of* $\mathbf{P}_{\{3\}\odot\{2\}}$ *contains information about* $\mathbf{f}^{Inv}(\mathbf{f}(\mathbf{P}_3))^{-1}$. *Therefore, in case of QPSK or higher PSK modulation, user 2 will have to wait for more retransmissions.*

We now look at the complexity of finding the minimum number of error-free retransmissions required by QAIDNC. Let $\mathbf{\Phi}^o$ be the K-by-N matrix such that $\mathbf{\Phi}^o(i,r) = 1$ if $\mathbf{\Phi}(i,r) \geq T$ and $\mathbf{\Phi}^o(i,r) = 0$ if $\mathbf{\Phi}(i,r) < T$. Therefore, $\mathbf{\Phi}^o$ describes the erasure patterns of the packets in the initial phase in the IDNC settings. It was shown in [16] that the problem of finding the minimum number of error-free retransmissions required by IDNC is NP-hard. To see the complexity of finding the minimum number of error-free retransmissions required by QAIDNC, one observes that it is a generalized version of the same problem in IDNC, as $\mathbf{\Phi}^o$ can be treated as a special realization of $\mathbf{\Phi}$. This implies that even a special case of finding the minimum number of error-free retransmissions required by QAIDNC is NP-hard, implying that in general the problem is NP-hard.

The detailed steps of QAIDNC followed by further explanations are given below:

[QA-1.] If there is an element in $\mathbf{\Phi}$ that is smaller than T, goto the next step. Otherwise stop.

[QA-2.] Let $\mathbf{\Phi}(i,*)$ be the ith row and $\mathbf{\Phi}(*,r)$ be the rth column of $\mathbf{\Phi}$, respectively. Rearrange the columns in $\mathbf{\Phi}$ and get $\mathbf{\Phi}' = [\mathbf{\Phi}(*,a_1)\ \mathbf{\Phi}(*,a_2)\ \cdots \mathbf{\Phi}(*,a_N)]$, where $\mathbf{\Phi}(*,a_i)$ is the column in $\mathbf{\Phi}$ that has the i-th largest number of elements whose values are smaller than T. In case $\mathbf{\Phi}(*,a_m)$ and $\mathbf{\Phi}(*,a_n)$ have the same number of elements that are less than T, the one with the larger sum of elements which are less than T will be placed before the other one.

[QA-3.] Let $\mathcal{I} = \{a_1\}$. From $i = 2$ to $i = N$, where there exists at least an element in $\mathbf{\Phi}(*,i)$ that is less than T, put a_i into $\mathcal{I}$ if a network coded packet formed by source packets with indices in $\{i\} \cup \mathcal{I}$ satisfies C.1-3 for all users.

[QA-4.] If $|\mathcal{I}| > 1$, retransmit the network coded packet formed by source packets with indices in $\mathcal{I}$. Otherwise retransmit $\mathbf{P}_{a_1}$. Update $\mathbf{\Phi}$ and go back to step 1.

We now demonstrate how $\mathcal{I}$ and the packet to be retransmitted are determined, when BPSK and higher PSK modulation

schemes are used, respectively.

A. BPSK Case

Let $\mathcal{I} = \{a_1,\ a_2,\ ...,\ a_n\}$. Also let $\mathcal{I}_k \triangleq \{a_i : a_i \in I,\ \mathbf{\Phi}(k, a_i) < T\}$. It is easy to see that the network coded packet $\mathbf{P}_{a_1} + \mathbf{P}_{a_2} + \cdots + \mathbf{P}_{a_n}$ satisfies:

(S.1) C.1 for each user k, if $\mathcal{I}_k = \{a_{r_1},\ a_{r_2}\}$ and $\mathbf{\Phi}(k, a_{r_1}) + \mathbf{\Phi}(k, a_{r_2}) \geq T$.

(S.2) C.2 for each user k, if $|\mathcal{I}_k| = 1$.

(S.3) C.3 for each user k, if $|\mathcal{I}_k| = 0$.

To check whether a_i should be included in $\mathcal{I}$ or not, I_k should be updated and S.1-3 should be examined, where $k \in [1, K]$.

For the convenience of future discussion, we define the K-dimension vector $\mathbf{c}_1$, where given $\mathbf{\Phi}$ and $\mathcal{I}$, $\mathbf{c}_1(k) = 1$ if S.1 is true for user k and $\mathbf{c}_1(k) = 0$ if S.1 is false for user k.

B. Higher PSK Case

It has been shown in Example 2 that when higher PSK modulations are used, a network coded packet formed by source packets with indices in $\mathcal{I}$ may not be 2-instantly decodable to every user k, where $\mathbf{c}_1(k) = 1$. In the following, we reveal this fact and provide extra steps upon the checking procedures of BPSK case which will make sure the source finds retransmission packets that satisfy C.1-3 when high PSK modulations are used.

Given $\mathbf{\Phi}$ and $\mathcal{I}$, let $\mathbf{c}_1$ has non-zero entries. To determine whether there is a packet formed by a combination of source packets with indices in $\mathcal{I}$ that is 2-instantly decodable to each user k, where $\mathbf{c}_1(k) = 1$, we can draw a *conflict graph* by the following procedure:

Algorithm 1 Conflict Graph Λ

Require: $\mathbf{\Phi}$, $\mathcal{I} = \{a_1, a_2, ..., a_n\}$, $\mathbf{c}_1$, T

1: **for** $k = 1$ **to** K **do**
2: **if** $\mathbf{c}_1(k) = 1$ **then**
3: Find a_i and a_r such that $a_i \in \mathcal{I}$, $a_r \in \mathcal{I}$, $a_i \neq a_r$, $\mathbf{\Phi}(k, a_i) < T$, $\mathbf{\Phi}(k, a_r) < T$
4: Create nodes named a_i and a_r if no node with either name has been created. Connect node a_i with a_r if they are not connected
5: **end if**
6: **end for**

After drawing a conflict graph, it is not difficult to see that if node a_r and node a_j are connected, then there exists $\mathcal{K}' \subseteq \mathcal{K}$, such that the users in $\mathcal{K}'$ has not decoded $\mathbf{P}_{a_i}$ and $\mathbf{P}_{a_r}$. Also, for $k \in \mathcal{K}'$, $\mathbf{\Phi}(k, a_i) + \mathbf{\Phi}(k, a_r) \geq T$. Moreover, there is no node in the graph that has degree 0 (i.e., all nodes are adjacent to at least one edge).

We now show in the following proposition that when a PSK modulation scheme higher than BPSK is used, under what condition can we find a network coded packet formed by source packets having indices in $\mathcal{I}$ which is 2-instantly decodable to every user k, where $\mathbf{c}_1(k) = 1$.

Proposition 2. *Let the source apply M-ary PSK modulation scheme, where $M = 2^\alpha$, $\alpha \in \mathbb{Z}^+ \setminus \{1\}$. Given $\mathbf{\Phi}$, $\mathcal{I}$ and $\mathbf{c}_1$, draw the conflict graph Λ by Algorithm 1. There exists a network coded packet $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}} \in \mathcal{Q}^{\mathrm{QANC}}$, where $\mathcal{A} \cup \mathcal{B} = \mathcal{I}$, which is 2-instantly decodable to every user k for which $\mathbf{c}_1(k) = 1$, if and only if Λ is bipartite.*

Proof: We will first show that if a conflict graph is not bipartite, we cannot find a network coded packet formed by source packets with indices in $\mathcal{I}$ that is 2-instantly decodable to every user k, where $\mathbf{c}_1(k) = 1$. Then, we demonstrate that we can find such a network coded packet when the conflict graph is bipartite.

Consider a conflict graph Θ with the set of nodes in $\mathcal{I}$. Let Θ be a non-bipartite graph, so that there exists an odd cycle $a_1 - a_2 - \cdots - a_n - a_1$, where n is an odd number and $\{a_1, a_2, ..., a_n\} \in \mathcal{I}$. For $i = 1, 2, \ldots, n$, define

$$\mathcal{K}_i \triangleq \{k \in \mathcal{K} : \mathbf{\Phi}(k, a_i) < T, \mathbf{\Phi}(k, a_{(i \mod n)+1}) < T \text{ and } \mathbf{\Phi}(k, a_i) + \mathbf{\Phi}(k, a_{(i \mod n)+1}) \geq T\}.$$

We now wish to find $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ which is 2-instantly decodable to the users in $\bigcup_{i\in[1,n]} \mathcal{K}_i$, where $\mathcal{A} \cup \mathcal{B} = \{a_1, a_2, ..., a_n\}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$. Since n is an odd number, there always exists $i \in [1, n]$ such that a_i and $a_{(i \mod n)+1}$ are both in either $\mathcal{A}$ or $\mathcal{B}$. Then $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ will not be 2-instantly decodable to users in $\mathcal{K}_i$. Therefore, it is not possible to find $\mathcal{A}' \subset \mathcal{I}$, $\mathcal{B}' \subset \mathcal{I}$, and $\mathcal{A}' \cup \mathcal{B}' = \mathcal{I}$, such that $\mathbf{P}_{\mathcal{A}'\odot\mathcal{B}'}$ is 2-instantly decodable to all users in $\bigcup_{i\in[1,n]} \mathcal{K}_i$.

Let Λ be a bipartite graph consists of two sets of nodes, namely $\mathcal{A} = \{a_1,\ a_2,\ ...,\ a_m\}$ and $\mathcal{B} = \{b_1,\ b_2,\ ...,\ a_n\}$, where nodes from $\mathcal{A}$ are only adjacent to nodes from $\mathcal{B}$ and vice versa, $\mathcal{A} \cup \mathcal{B} = \mathcal{I}$. Let $a_i \in \mathcal{A}$ is connected to nodes in $\mathcal{B}_{a_i}$, where $\mathcal{B}_{a_i} \subseteq \mathcal{B}$. Also, define

$$\mathcal{K}'_{a_i} \triangleq \{k \in \mathcal{K} : \mathbf{\Phi}(k, a_i) < T,\ \text{and } \exists b_k \in \mathcal{B}_{a_i} \text{ such that } \mathbf{\Phi}(k, b_k) < T \text{ and } \mathbf{\Phi}(k, a_i) + \mathbf{\Phi}(k, b_k) \geq T\},$$

where by the above definition, $\mathbf{c}_1(k) = 1$ if $k \in \mathcal{K}'_{a_i}$. Since user k in $\mathcal{K}_{a_i}$ has decoded packets with indices in $\mathcal{A} \setminus \{a_i\}$ and $\mathcal{B} \setminus \{b_k\}$, from Proposition 1, upon receiving $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$, the user can decode $\mathbf{P}_{a_i}$ and $\mathbf{P}_{b_r}$. Therefore, if Λ is a bipartite graph, the packet $\mathbf{P}_{\mathcal{A}\odot\mathcal{B}}$ is 2-instantly decodable to every user k, where $\mathbf{c}_1(k) = 1$. ■

Armed with the above proposition, we can conclude that given $\mathbf{\Phi}$, $\mathcal{I}$, a_i and $\mathbf{c}_1$ with respect to $\mathcal{I} = \mathcal{I} \cup \{a_i\}$, if the conflict graph drawn according to Algorithm 1 is not bipartite, then we should not include a_i into $\mathcal{I}$.

C. Complexity of QAIDNC

We now discuss the encoding and decoding complexity of QAIDNC.

For BPSK case: It takes $O(KN)$ operations to find out how many elements in each column of $\mathbf{\Phi}$ have their values less than T, and then $O(N \log N)$ operations to do the sorting. Therefore, the total encoding complexity is $O(KN + N \log N)$.

For higher PSK cases: Some extra work needs to be done thanks to the conflict graph. Note that there can be at most K edges and $2K$ nodes in a conflict graph, and it takes $O(K)$ operations to check whether the graph is bipartite. Therefore, the total encoding complexity remains unchanged, namely $O(KN + N \log N)$.

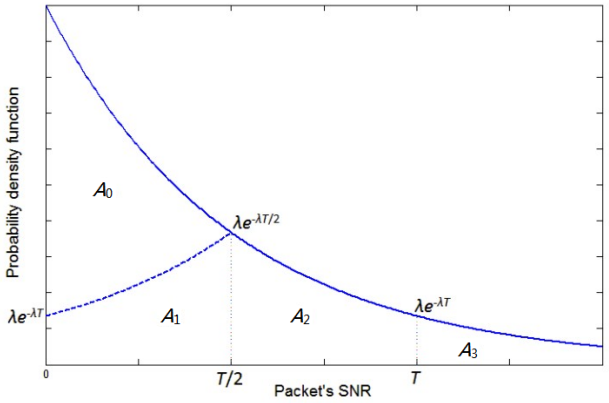


Fig. 2. Distribution of the SNR of a received packet at the user side. Since the channel is subject to Rayleigh fading, the SNR follows exponential distribution, where the probability density function is given by $\lambda e^{-\lambda x}$, $x \geq 0$ with x being the random variable.

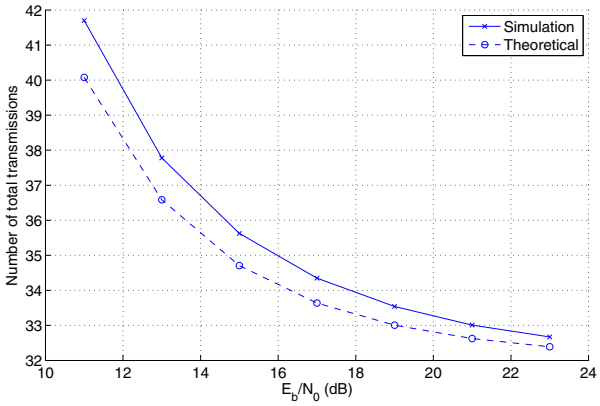


Fig. 3. QAIDNC single user case, where $N = 32$. The channel is subject to Rayleigh fading.

TABLE I
VERIFYING THE ACCURACY OF THE ANALYSIS

N	8	64	512	4096
$\frac{R_{total}-R_{Sim}}{R_{total}}$	8.9%	2.7%	0.7%	0.2%

As a retransmission packet can be 2-instantly decodable, instantly decodable, or worthless to a user, the complexity of decoding a packet received in the retransmission phase using QAIDNC is $O(N)$.

D. Analyzing QAIDNC

In this section, we give an asymptotic analysis on the average number of total transmissions required by QAIDNC when there is a single user to serve, and the user will not utilize erroneous retransmission packets as described from M-1 to M-3.

Let h be the channel gain of a particular transmission, and assume the source transmits packets with power 1. Since $|h|$ follows Rayleigh distribution, the SNR of a packet at the user side, namely $\frac{|h|^2}{\sigma^2}$, follows exponential distribution with parameter $2\sigma^2$ [29]. For simplicity let $\lambda = 2\sigma^2$. A graph showing the shape of an exponentially distributed random variable is shown in Fig. 2.

Let N be sufficiently large. From Fig. 2, we can see that after the initial phase, there will be on average $N \cdot A_3$ source packets that have their SNRs larger than or equal to T, where A_3 is the area beneath the density curve and to the right of $x = T$. Also, the source will find on average $N \cdot A_2$ coded packets that are 2-instantly decodable to the user, where A_2 is the area beneath the density curve and bounded by $x = T/2$ and $x = T$ from the left and the right, respectively. This is because the source can retransmit $\mathbf{P}_{\{m\}\odot\{n\}}$, where $\mathbf{P}_m$ and $\mathbf{P}_n$ has SNR $T/2-\Delta$ and $T/2+\Delta$, respectively, $\Delta \in [0, T/2]$, so that the user can decode $\mathbf{P}_m$ and $\mathbf{P}_n$ once she decodes $\mathbf{P}_{\{m\}\odot\{n\}}$. Finally, the source will have to retransmit each of the remaining $N \cdot A_0$ source packets, where A_1 is the mirror of A_2 with respect to $x = T/2$ and A_0 is the area beneath the density curve apart from A_1, A_2, and A_T. The areas A_3, A_2, A_1, and A_0 can be calculated as follows:

$$A_3 = \int_T^\infty \lambda e^{-\lambda x} dx = [-e^{-\lambda x}]_T^\infty = e^{-\lambda T},$$

$$A_2 = \int_{T/2}^T \lambda e^{-\lambda x} dx = [-e^{-\lambda x}]_{T/2}^T = e^{-\frac{\lambda T}{2}} - e^{-\lambda T},$$

$$A_2 = A_1,$$

$$A_0 = \int_0^{T/2} \lambda e^{-\lambda x} dx - A_1 = [-e^{-\lambda x}]_0^{T/2} - A_1 = 1 - e^{-\frac{\lambda T}{2}}.$$

To find out the average number of transmissions before the user can receive a packet correctly, namely $R_{correct}$, notice that each transmission has a probability of A_T to be successful, and therefore

$$R_{correct} = A_3 + 2(1-A_3)A_3 + 3(1-A_3)^2 A_3 + \cdots + n(1-A_3)^{n-1} A_3 + \cdots.$$

Observe that $0 < A_3 < 1$, and

$$\begin{aligned} R_{correct} - (1-A_3)R_{correct} &= A_3[1 + (1-A_3) + (1-A_3)^2 + \cdots + (1-A_3)^n + \cdots] \\ &= A_3 \tfrac{1}{1-(1-A_3)} = 1, \end{aligned}$$

so

$$R_{correct} = \frac{1}{A_3} = e^{\lambda T}.$$

Therefore, the average total number of transmissions R_{total} can be calculated as

$$R_{total} = N + R_{correct} N(A_0 + A_2) = N e^{\lambda T} = N e^{2\sigma^2 T}.$$

To verify the above analysis, Fig. 3 shows the simulation results and the analytical results, where $N = 32$ and the channel is subject to Rayleigh fading. The figure shows that the analysis has an error of only about 3% even if N is relatively small. Table I shows the difference of R_{total} and the simulation results of the total number transmissions denoted as R_{Sim} in percentage, where $E_b/N_0 = 10$ dB. We can see that the analytical result becomes more accurate as N increases.

VI. REFERENCE SCHEMES

Before going into simulations, we describe some details about the reference schemes that will be compared to our proposed scheme.

A. CC-HARQ

The CC-HARQ scheme is simple: for each source packet, if there is any user who has not received it correctly, repeatedly retransmit that source packet until all users have received that packet correctly. A user will use MRC to combine all the copies of a source packet it receives.

B. IDNC

In IDNC, a retransmission packet is required to be instantly decodable or worthless to each user. In the simulations, we apply the algorithm in [16] to evaluate the performance of IDNC. A user will simply throw away an erroneous received packet in the retransmission phase.

C. IDNC with Memory

To make a more fair comparison with QAIDNC, we modify the original design of IDNC, by allowing the users to utilize erroneous retransmission packets in the way described from M-1 to M-3.

D. RLNC

In each round of retransmission, the source will generate a network coded packet which is a linear combination of the N source packets, represented as $\mathcal{Q}^{\mathrm{LNC}}$, and retransmit it once regardless of whether it is received correctly by any of the users. Specifically, the coefficients β_i, $i \in \mathcal{N}$ are drawn randomly from $GF(q)$. Although it is possible that RLNC may generate packets that are not innovative to every user even if $q \geq K$, we assume that it can always generate innovative packets, showing its idealized performance.

E. SYNC

Let $\mathbf{P}_n$ and $\mathbf{P}_{n+1}$ be 2 packets each contains $\zeta\eta$ bits. $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ returns a packet with $2\zeta\eta$ bits, with the odd bits coming from $\mathbf{P}_n$ in the same order and the even bits coming from $\mathbf{P}_{n+1}$ in the same order. A network coded retransmission scheme designed for single source single user, namely SYNC, is proposed in [22]. The basic procedure of SYNC goes like this: First transmit $\mathbf{P}_n$ using m-PSK. In case $\mathbf{P}_n$ cannot be correctly decoded, send $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ using $2m$-PSK and try to decode $\mathbf{P}_n$. If $\mathbf{P}_n$ is still not correctly received, repeatedly send $\mathbf{P}_n$ using m-PSK until it can be correctly decoded.

The idea behind SYNC is that, since the source will transmit $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ if the first transmission of $\mathbf{P}_n$ fails, the user will have an observation of $\mathbf{P}_{n+1}$ after it can decode $\mathbf{P}_n$. We refer to interested readers to [22] for more details.

Fig. 4 plots the BER of $\mathbf{P}_n$ using the SYNC scheme after the source sends $\mathbf{P}_n$ once and $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$. The source uses BPSK modulation when sending $\mathbf{P}_n$ and QPSK modulation when sending $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$, respectively. The BER of $\mathbf{P}_n$ using SYNC after sending $\mathbf{P}_n$ and $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ is indeed decreased, and the curve has about 2 dB gain compared to the curve of BPSK BER (i.e., send $\mathbf{P}_n$ once using BPSK).

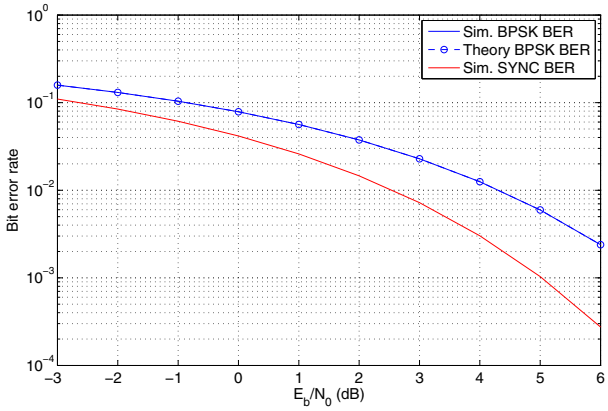


Fig. 4. BER of $\mathbf{P}_n$ after transmitting $\mathbf{P}_n$ once using BPSK and $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ using QPSK, where each of $\mathbf{P}_n$ and $\mathbf{P}_{n+1}$ contains 12000 bits. The BER of BPSK modulation is plotted for reference. "Sim." in the legend means simulation. This graph is a reproduction of Fig.8 in [22], with slight difference on the x-axis.

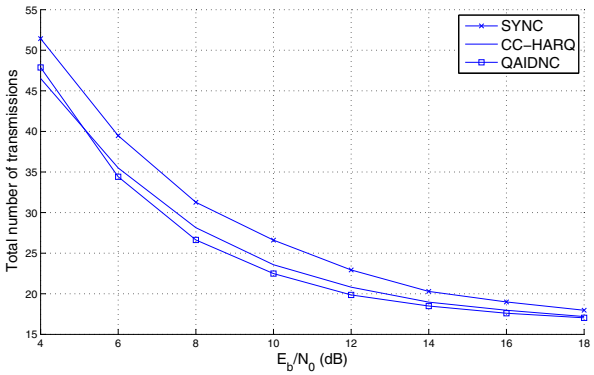


Fig. 5. Comparison among SYNC, CC-HARQ, and QAIDNC, where there is only one user to serve. The source is to deliver 16 packets, each with 12000 bits. BPSK is used for CC-HARQ and QAIDNC. When sending a source packet, SYNC also uses BPSK.

VII. Simulation Results

In this section, we compare the performance of QAIDNC with SYNC, IDNC and RLNC schemes. As a reference we also plot the performance of CC-HARQ. In all subsequent simulations, we treat a packet as being correctly received if the BER of the packet after detection is less than or equal to $\varepsilon_{th} = 10^{-3}$. Given the BER threshold ε_{th}, the corresponding SNR threshold T when the source uses BPSK and QPSK can be calculated as $\varepsilon_{th} = \frac{1}{2}\mathrm{erfc}(\sqrt{T})$ and $\varepsilon_{th} = \mathrm{erfc}(\sqrt{\frac{T}{2}})$, respectively, where

$$\mathrm{erfc} \triangleq \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2) dt.$$

Fig. 5 shows the total number of transmissions required for the source to deliver 16 packets to the single user. The channel between the source and the user is subject to Rayleigh fading. We can observe that when the power per bit over noise spectral density (E_b/N_0) is low, QAIDNC could perform worse than CC-HARQ. At high E_b/N_0, however, QAIDNC performs the best. This is because it is more likely to find pairs of packets who have their sum SNRs larger than T, and at the same time

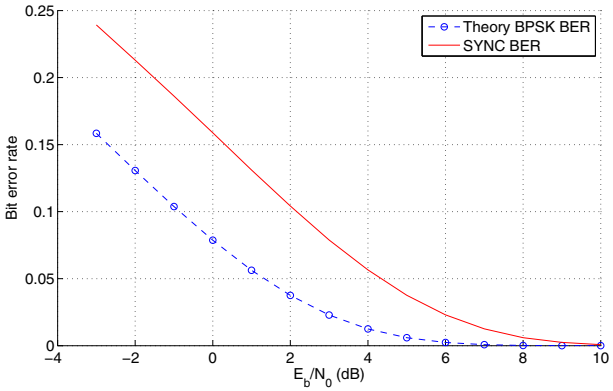


Fig. 6. The BER of $\mathbf{P}_{n+1}$, with an observation of $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ corrupted by AWGN only (i.e. no channel fading) and the correct version of $\mathbf{P}_n$, shown in red curve. $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ is transmitted using QPSK. The BER of BPSK modulation over AWGN channel is also plotted for reference.

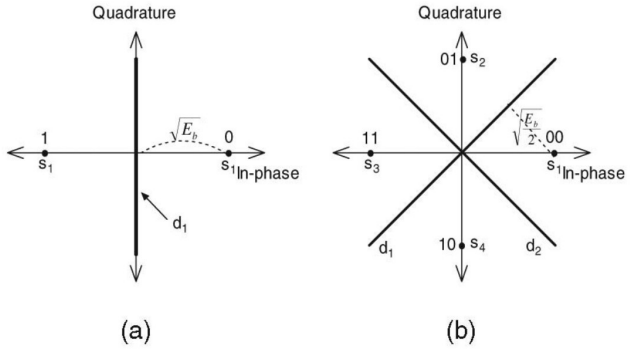


Fig. 7. The constellations of PSK modulation schemes. (a) BPSK, with d_1 as the decision boundary. (b) QPSK, with d_1 and d_2 as the decision boundaries. This graph is taken from Fig. 3 in [22].

it is more likely that a retransmission packet can be correctly decoded. We can also observe that SYNC always has poorer performance than CC-HARQ. To find out the reason, we plot the BER of $\mathbf{P}_{n+1}$, when the user has an observation of $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ affected by AWGN only and the correct version of $\mathbf{P}_n$ in Fig. 6. We can see that the BER of $\mathbf{P}_{n+1}$ is actually about 3 dB worse than the BER of BPSK. To justify the 3 dB loss, let's temporally assume that each of $\mathbf{P}_n$ and $\mathbf{P}_{n+1}$ has one bit only and $\mathbf{P}_n = 0$. In this case, $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ can only be either 00 or 01. With the knowledge of $\mathbf{P}_n$, the detection of $\mathbf{P}_{n+1}$ is effectively reduced to BPSK detection, with the distance from the decision boundary (d_1) to either of the symbols as $\sqrt{\frac{E_b}{2}}$. Fixing the power per symbol, the aforementioned distance is $\sqrt{\frac{1}{2}}$ of the distance from the decision boundary to the symbols if BPSK is applied to send $\mathbf{P}_{n+1}$. In other words, the effective SNR of trying to decode $\mathbf{P}_{n+1}$ from a correct version of $\mathbf{P}_n$ and a corrupted QPSK modulated $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ is half (about -3 dB) of the SNR of just sending $\mathbf{P}_{n+1}$ using BPSK. Together with the observation of Fig. 4, we can see that although $\mathbf{P}_n \uplus \mathbf{P}_{n+1}$ can help improving the BER of $\mathbf{P}_n$ by about 2 dB, the BER of $\mathbf{P}_{n+1}$ after correctly decoding $\mathbf{P}_n$ is 3 dB lower than just transmitting $\mathbf{P}_{n+1}$ using BPSK, ending up with a net result of about 1 dB loss compared to CC-HARQ.

In the following simulations, we will switch to multi-user cases. Besides Rayleigh fading, the channel gain of a source-to-user link is also affected by path loss. We assume that users are uniformly distributed in a circle with radius 1. In the simulations, we incorporate path loss using a simplistic model, where the path loss of a source-to-user link is $\frac{1}{d^3}$ as small as the path loss of the user located on the boundary, and d is the distance between the center of the circle and the user. We also force d to be 0.1 if $d < 0.1$ to avoid the unrealistic situations that the received power of a user near the transmitter can be infinitely larger than that of a user located at the boundary. The SYNC scheme will not be considered in subsequent simulations, as it has poorer performance than CC-HARQ as previously demonstrated and [22] does not give a broadcast algorithm for SYNC.

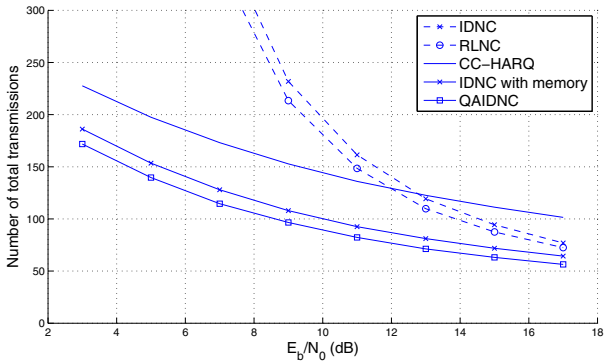
Fig. 8 shows the performance of CC-HARQ, IDNC, RLNC, IDNC with memory, and QAIDNC when the source uses BPSK, where $K = 64$ and $N = 32$. From Fig. 8(a), we can observe that both IDNC and RLNC suffer from very bad performance when E_b/N_0 is relatively low, due to the reason that they throw away any erroneous packets which are frequent in the retransmission phase. On the other hand, IDNC with memory and QAIDNC overcome this problem as they keep retransmitted packets for future use. Also, equipped with the knowledge of the SNRs of the source packets of the users, QAIDNC can find more coding opportunities than IDNC with memory and about 10% improvement on total number of transmissions (i.e. the sum of number of transmissions in the initial phase and in the retransmission phase) is achieved by QAIDNC compared to IDNC with memory. Although QAIDNC may need more space to store the coded packets in the retransmission phase than IDNC with memory, as observed in Fig. 8(a), the memory requirement is far from too much. Specifically, the maximum number of stored coded packets by QAIDNC when $E_b/N_0 = 3$ dB is about the same as N, and this number drops quickly as E_b/N_0 increases.

Fig. 8(c) plots the Cumulative Distribution Function (CDF) of the users' average decoding delay when different retransmission schemes are used at $E_b/N_0 = 7$ dB. Let l_n be the number of source packets that can be decoded by a particular user after the nth retransmission. The average decoding delay of the user is calculated as

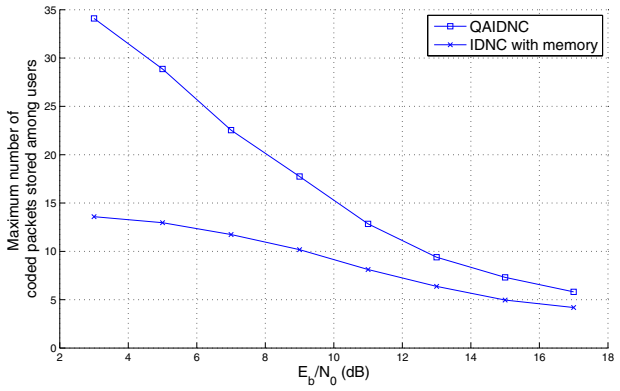
$$\frac{\sum_{n=1}^{R_{retran}} n \cdot l_n}{\sum_{n=1}^{R_{retran}} l_n},$$

where R_{retran} is the number of total retransmissions before the user correctly decodes all N source packets. It can be seen that QAIDNC has the lowest average decoding delay. RLNC, on the other hand, has 4-6 fold average decoding delay as that of QAIDNC.

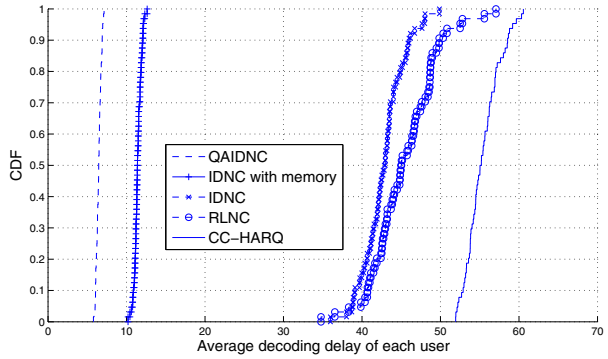
Fig. 9 shows a similar simulation, where $K = 64$ and $N = 32$, but the source now applies QPSK. It can be observed from Fig. 9(a) that QAIDNC has a better improvement than IDNC with memory at low E_b/N_0, more than 20% decrease on total number of transmissions when $E_b/N_0 = 5$ dB. From Fig. 9(b), we can see that the memory requirement of QAIDNC when QPSK is used can be higher than the situation when BPSK is used. There are two reasons: (1) T is higher when QPSK is used compared to when BPSK is used, and therefore more retransmissions will occur in the QPSK case.



(a)

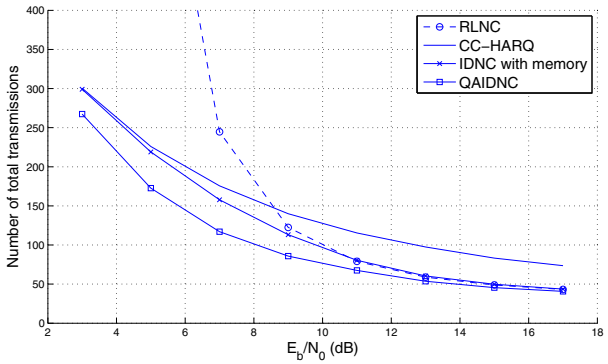


(b)

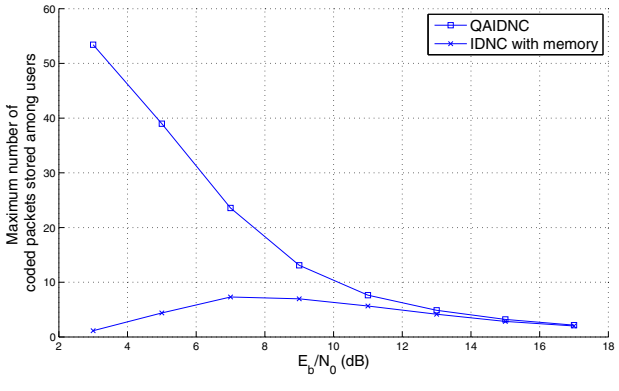


(c)

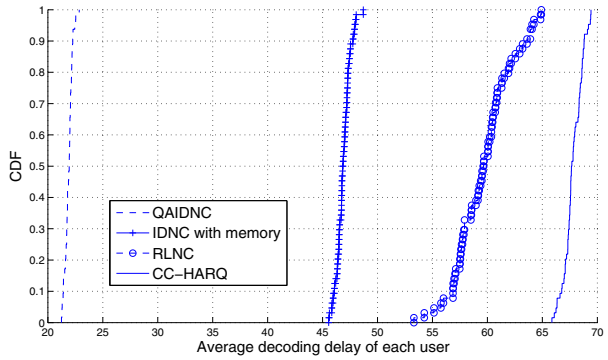
Fig. 8. Evaluate performances of various schemes when the source uses BPSK, where $K = 64$ and $N = 32$. (a) Number of total transmissions. (b) Memory requirement to store the coded packets for QAIDNC and IDNC with memeory. (c) Decoding delay when $E_b/N_0 = 7$ dB.



(a)



(b)



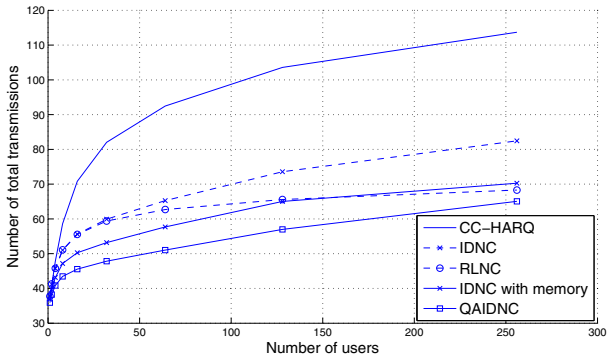
(c)

Fig. 9. Evaluate performances of various schemes when the source uses QPSK, where $K = 64$ and $N = 32$. (a) Number of total transmissions. (b) Memory requirement to store the coded packets for QAIDNC and IDNC with memeory. (c) Decoding delay when $E_b/N_0 = 7$ dB.

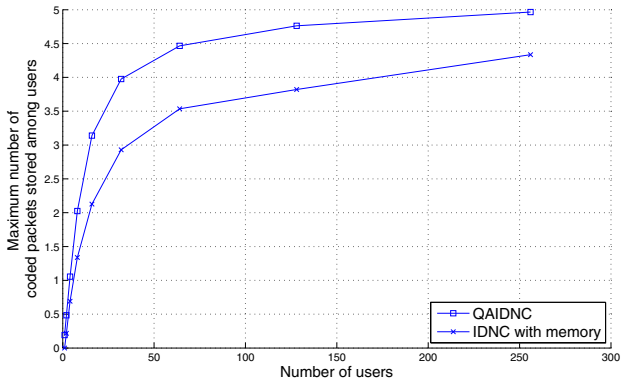
(2) As explained in Example 3, the stored coded packets are not so conveniently used in QPSK case compared to BPSK case. Also, IDNC with memory uses very little memory when E_b/N_0 is low, which indicates that with high T and low transmission power, it is difficult for IDNC with memory to find coding opportunities. The fact that IDNC with memory performs almost the same as CC-HARQ when $E_b/N_0 = 3$ dB also reflects this phenomenon.

Fig. 10 shows the performance of the schemes when the number of users in the system K increases, where $E_b/N_0 = 9$ dB, $N = 32$, and BPSK is used. From Fig. 10(a) we can see that QAIDNC performs the best even if K is as large as about 250. When K is about 130, QAIDNC has an improvement of about 30% on the total number of transmissions compared to RLNC and IDNC with memory. Also, as demonstrated in Fig. 10(b), the memory requirement of QAIDNC grows very slowly as K grows. Fig. 11 shows the the same performance metrics, where where $E_b/N_0 = 12$ dB, $N = 32$, and QPSK is used. It can be observed that if K becomes even larger, RLNC would outperform QAIDNC. This is because it is becoming more and more difficult for QAIDNC to find a coded packet that satisfy C.1-C.3 as K is growing.

Fig. 12 simulates the performance of QAIDNC when the users' feedback is not guaranteed to reach the source. There

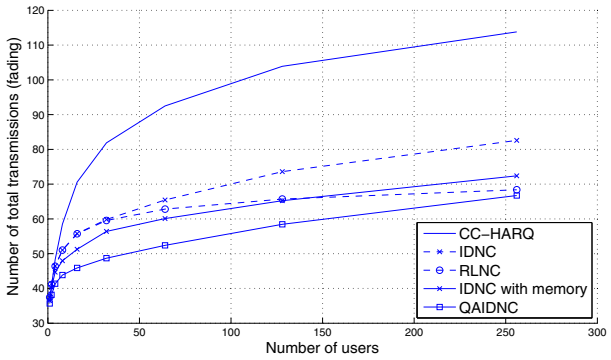


(a)

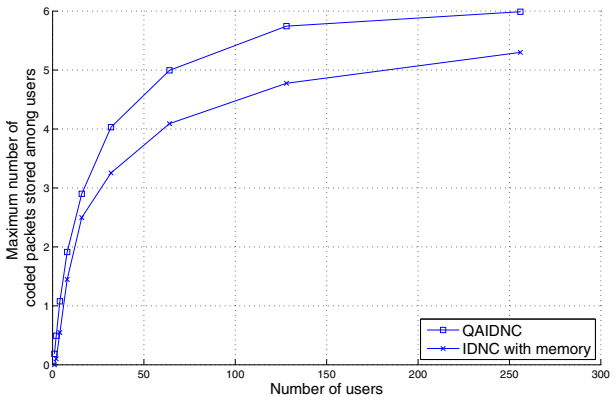


(b)

Fig. 10. Evaluate different schemes when the number of users changes, where $E_b/N_0 = 9$ dB and $N = 32$. BPSK modulation is used by the source. (a) Number of total transmissions. (b) Memory requirement to store the coded packets.



(a)



(b)

Fig. 11. Evaluate different schemes when the number of users changes, where $E_b/N_0 = 12$ dB and $N = 32$. QPSK modulation is used by the source. (a) Number of total transmissions. (b) Memory requirement to store the coded packets.

are 32 source packets to send to the 64 users, using BPSK. Each user has a probability of 0.3 of failing to inform the source her most updated source packets' SNRs when sending feedback. When a user's feedback gets to the source successfully, the source will know the SNRs of the source packets at the time the user sends the feedback. From the figure, we can observe that the impact of feedback loss is quite negligible when E_b/N_0 is relatively low. At high E_b/N_0 levels, the feedback loss effect becomes observable - about 10% increase on total number of transmissions when $E_b/N_0 = 23$ dB. The more obvious performance loss due to lossy feedback at high E_b/N_0 levels is due to the reason that the retransmissions are almost always successful, and therefore the impact of feedback loss stands out.

Fig. 13 compares the performance of QAIDNC when the source only knows the quantized versions of the SNRs of the source packets received by the users versus when the source knows exactly what the SNRs of the source packets at the users are. The feedback from the users is assumed to be always able to reach the source, and a user applies either 3-bit or 4-bit uniform quantization to the SNRs. The simulation shows that when $E_b/N_0 = 3$ dB, the source will have to transmit about 6% and 15% more packets when the SNRs in the feedback are quantized using 4 bits and 3 bits, respectively. As the E_b/N_0 level increases, the loss due to quantization becomes even less. For example, at $E_b/N_0 = 9$ dB, the loss due to quantization for 4-bit quantization and 3-bit quantization is about 3% and 7%, respectively.

VIII. Conclusion

In this work, we develop the idea of Quality-Aware Network Coding (QANC) based on Phase-Shift Keying (PSK) modulation schemes. We apply QANC to wireless broadcast channels and design Quality-Aware Instantly Decodable Network Coding (QAIDNC) schemes which outperforms Instantly Decodable Network Coding (IDNC) in terms of total number of transmissions and average decoding delay, while the encoding and decoding complexity of QAIDNC remain low. Also, QAIDNC outperforms Random Linear Network Coding when the number of users in the system is relatively small, suggesting that QAIDNC offers a good tradeoff between throughput and complexity. It is also verified by simulations that QAIDNC is quite robust against ACK loss.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions on improving the paper.

References

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

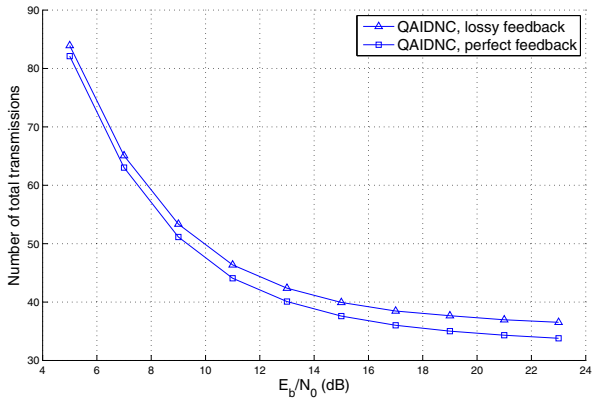


Fig. 12. Evaluating the impact of feedback loss to QAIDNC, $K = 64$, $N = 32$. Each user has a probability of 0.3 of failing to send her feedback to the source. BPSK modulation is applied.

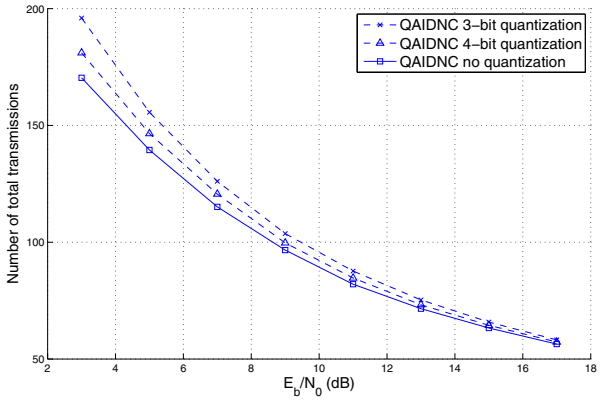


Fig. 13. Evaluating the impact of quantizating the SNRs of source packets in feedback to QAIDNC, $K = 64$, $N = 32$. The quantization is uniform. BPSK modulation is applied.

[2] M. Wang and B. Li, "R2: random push with random network coding in live peer-to-peer streaming," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 9, pp. 1655–1666, 2007.

[3] S. Omiwade, R. Zheng, and C. Hua, "Practical localized network coding in wireless mesh networks," in *Proc. 2008 IEEE Commun. Soc. Conf. Sensor, Mesh Ad Hoc Commun. Netw.*, pp. 332–340.

[4] C. Khirallah, D. Vukobratovic, and J. Thompson, "Performance analysis and energy efficiency of random network coding in LTE-advanced," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4275–4285, 2012.

[5] D. Nguyen, T. Tran, T. Nguyen, and B. Bose, "Wireless broadcast using network coding," *IEEE Trans. Veh. Technol.*, vol. 58, no. 2, pp. 914–925, 2009.

[6] L. Li, R. Ramjee, M. Buddhikot, and S. Miller, "Network coding-based broadcast in mobile ad-hoc networks," in *Proc. 2007 IEEE Int. Conf. Comput. Commun.*, pp. 1739–1747.

[7] P. Sadeghi and M. Yu, "Instantly decodable versus random linear network coding: a comparative framework for throughput and decoding delay performance," *CoRR*, vol. abs/1208.2387, 2012.

[8] R. Joda and F. Lahouti, "Network code design for orthogonal two-hop network with broadcasting relay: a joint source-channel-network coding approach," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 132–142, 2012.

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

[10] R. Ahlswede, N. Cai, S. Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, July 2000.

[11] C. Feng and B. Li, "On large-scale peer-to-peer streaming systems with network coding," in *Proc. 2008 ACM Int. Conf. Multimedia*, pp. 269–278.

[12] Q. Zhang, F. Fitzek, and V. Iversen, "Design and performance evaluation of cooperative retransmission scheme for reliable multicast services in cellular controlled p2p networks," in *Proc. 2007 IEEE Int. Symp. Pers., Indoor Mobile Radio Commun.*, pp. 1–5.

[13] M. Wang and B. Li, "How practical is network coding?" in *Proc. 2006 IEEE Int. Workshop Quality of Service*, pp. 274–278.

[14] H. Y. Kwan, K. Shum, and C. W. Sung, "Generation of innovative and sparse encoding vectors for broadcast systems with feedback," in *Proc. 2011 IEEE Int. Symp. Inf. Theory*, pp. 1161–1165.

[15] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard, and J. Crowcroft, "XORs in the air: practical wireless network coding," *IEEE/ACM Trans. Netw.*, vol. 16, no. 3, pp. 497–510, 2008.

[16] E. Rozner, A. P. Iyer, Y. Mehta, L. Qiu, and M. Jafry, "ER: efficient retransmission scheme for wireless LANs," in *Proc. 2007 ACM CoNEXT Conf.*, pp. 8:1–8:12.

[17] L. Keller, E. Drinea, and C. Fragouli, "Online broadcasting with network coding," in *Proc. 2008 Workshop Netw. Coding, Theory, Appl.*, pp. 68–73.

[18] S. Sorour and S. Valaee, "On minimizing broadcast completion delay for instantly decodable network coding," in *Proc. 2010 IEEE Int. Conf. Commun.*, pp. 1–5.

[19] ——, "Completion delay minimization for instantly decodable network codes," *CoRR*, vol. abs/1201.4768, 2012.

[20] P. Sadeghi, R. Shams, and D. Traskov, "An optimal adaptive network coding scheme for minimizing decoding delay in broadcast erasure channels," *Eurasip J. Wireless Commun. Netw.*, Apr. 2010.

[21] G. R. Woo, P. Kheradpour, D. Shen, and D. Katabi, "Beyond the bits: cooperative packet recovery using physical layer information," in *Proc. 2007 ACM Int. Conf. Mobile Comput. Netw.*, pp. 147–158.

[22] S. Yun, H. Kim, and K. Tan, "Towards zero retransmission overhead: a symbol level network coding approach to retransmission," *IEEE Trans. Mobile Comput.*, vol. 10, no. 8, pp. 1083–1095, 2011.

[23] Y. Liu and C. Sung, "Network-coded retransmissions in wireless demodulate-and-forward relay channels," *EURASIP J. Wireless Commun. Netw.*, vol. 2013, no. 1, p. 136, 2013.

[24] D. Chase, "Code combining—a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Trans. Commun.*, vol. 33, no. 5, pp. 385–393, May 1985.

[25] A. Eryilmaz, A. Ozdaglar, and M. Medard, "On delay performance gains from network coding," in *Proc. 2006 Conf. Inf. Sci. Syst.*, pp. 864–870.

[26] Y. Liu and C. W. Sung, "A cross-layer design of network coded retransmissions in wireless relay channels," in *Proc. 2011 IEEE Wireless Advanced Conf.*, pp. 201–206.

[27] T. K. Y. Lo, "Maximum ratio transmission," *IEEE Trans. Commun.*, vol. 47, no. 10, pp. 1458–1461, 1999.

[28] F. R. Kschischang, P. G. Debuda, and S. Pasupathy, "Block coset codes for M-ary phase-shift keying," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 900–913, Aug. 1989.

[29] V. Krishnan, *Probability and Random Processes (Wiley Survival Guides in Engineering and Science)*. John Wiley & Sons, Inc., 2006.

Ye Liu received the B.Eng. in electronic and communication engineering with first class honors from the City University of Hong Kong, Hong Kong SAR, in 2009. He is now pursuing his Ph.D. in electronic engineering at the same university. In the summer of 2013, he visited the Laboratory of Information, Networking and Communication Sciences (LINCS) located in Paris, France. His research interests include network coding and LTE system optimizations.

Chi Wan Sung (M'98) received his B.Eng, M.Phil, and Ph.D. degrees in information engineering from the Chinese University of Hong Kong in 1993, 1995, and 1998, respectively. He joined the faculty at City University of Hong Kong in 2000, and is now Associate Professor with the Department of Electronic Engineering. He is an Adjunct Associate Research Professor at the University of South Australia, and is on the editorial boards of the *Transactions on Emerging Telecommunications Technologies* (ETT) and the *ETRI Journal*. His research interests include wireless communications, network coding, cloud storage systems, and algorithms and complexity.