Extreme Value FEC for Wireless Data Broadcasting

Weiyao Xiao* and David Starobinski*
* Department of Electrical and Computer Engineering
Boston University, Boston, MA 02215
Email:{weiyao, staro}@bu.edu

Abstract*—The advent of practical rateless codes enables implementation of highly efficient packet-level forward error correction (FEC) strategies for reliable data broadcasting in loss-prone wireless networks. Yet, the critical question of accurately quantifying the proper amount of redundancy has remained largely unsolved. In this paper, we exploit advances in extreme value theory to rigorously address this problem. Under the asymptotic regime of a large number of receivers, we derive a closed-form expression for the cumulative distribution function (CDF) of the completion time of file distribution. We show the existence of a phase transition associated with this CDF and accurately locate the transition point. We derive tight convergence bounds demonstrating the accuracy of the asymptotic estimate for the practical case of a finite number of receivers. We also provide an asymptotic closed-form expression on the expected completion time under heterogeneous packet loss. We demonstrate the benefits of our approach through simulation and through real experiments on a testbed of 20 Tmote Sky sensors. Specifically, we augment the existing Rateless Deluge software dissemination protocol with an extreme value FEC strategy. The experimental results reveal reduction by a factor of five in retransmission request messages and by a factor of two in total dissemination time, at the cost of a marginally higher number of data packet transmissions in the order of 5%.

I. Introduction

Reliable data broadcasting for wireless networks is an essential service supporting a plethora of applications, including distribution of text and multimedia contents, podcasting, and over-the-air programming (OAP) [1–4].

The lossy nature of wireless channels significantly complicates the task of reliable broadcasting, however. Due to the potentially large number of wireless devices, the so-called "broadcast storm" [5] phenomenon may arise when multiple receivers contend over a shared channel to request retransmissions of lost packets via either acknowledgements (ACKs) or negative acknowledgements (NACKs) messages. Although some mechanisms exist to mitigate the broadcast storm problem, such as ACK and NACK suppression [6], the impact of this problem can still be considerable [2, 4].

Ideally, instead of relying on receivers to notify a source about missing packets, a procedure commonly referred to as automatic repeat request (ARQ), the source should be able to accurately predict the total number of transmissions required and sends out data without the need for acknowledgements. Packet-level forward error correction (FEC) [7] provides a practical approach towards implementing this idea. With the advent of rateless codes, such as random linear codes, LT, and raptor codes [8, 9], FEC can be implemented in a very efficient fashion, whereas a source continuously encodes new packets based on the M original packets of a given file. The source then sends out the encoded packets and as soon as receiver obtains M (or slightly more) distinct packets, it can reconstruct the entire file successfully.

Although FEC broadcasting has been shown to outperform ARQ in many cases [10, 11], the major issue of quantifying the proper amount of redundancy has remained unsolved to a large extent (cf. Section II for related work). While transmitting too many redundant packets wastes bandwidth and energy, too little redundancy leaves many receivers unable to reconstruct the original file, leading to retransmission requests and eventually the same problems as encountered by ARQ schemes.

In this paper, we exploit advances in extreme value theory (EVT) [12] to rigorously address the problem of quantifying FEC redundancy in lossy wireless broadcast networks. Our main contributions are as follows.

First, under the asymptotic regime of a large number of receivers N, we derive a *closed-form* expression for the cumulative distribution function (CDF) of the completion time (i.e., the total number of packets to be sent by a source to ensure file recovery by all the receivers). Our analysis reveals the existence of a *phase transition* property associated with this CDF. Specifically, we show that there exists a threshold on the number of packets to be sent below which the probability that a file can be recovered by all the nodes in the network is close to zero. However, if the number of packets sent is slightly greater than the threshold, then the probability that every node in the network is able to reconstruct the file quickly approaches one.

Our second contribution is the derivation of tight *convergence bounds* for a finite number of receivers N. These bounds allow us to estimate the error committed by replacing the exact CDF by its limiting form. They also provide a means to compute the amount of redundancy needed for finite values of N. The bounds reveal that the asymptotic formula is remarkably accurate even for small values of N (e.g., 10 or 20).

Third, we analyze the *heterogeneous* packet loss case, whereby different receivers experience different packet loss probabilities. To this effect, we establish a relationship between the data broadcasting problem and the multi-set coupon collector's problem [13]. Exploiting this relationship, we provide an asymptotic closed-form expression on the expected number of transmissions needed to successfully disseminate a file to a set of receivers with heterogeneous packet loss probabilities.

Last, we conduct *real experiments* on a testbed of 20 Tmote

This research was supported in part by the US National Science Foundation under grant numbers CNS-0132802, CNS-0435312, CCF-0729158.

Sky sensors that illustrate practical use of our theoretical findings. Specifically, we embed our extreme-value FEC strategy into the Rateless Deluge OAP protocol [4] and demonstrate potential for significant reduction in retransmission requests (about 80%) and in completion time (about 50%) at the cost of marginally higher data packet transmissions (less than 5%) with respect to the original protocol.

The rest of this paper is organized as follows. We first discuss related research on FEC data broadcasting in Section II. After reviewing basics of extreme value theory in Section III, we present our network model and problem formalization for homogeneous packet loss in Section IV-A, conduct an asymptotic analysis of FEC broadcasting as $N \to \infty$ in Section IV-B, and derive convergence bounds for finite N in Section IV-C. An asymptotic analysis of the expected completion time, under heterogeneous packet loss, is carried out in Section V. We present our simulation results and prototype implementation in Sections VI and VII respectively, and conclude the paper in Section VIII. Due to space limitation, we only provide sketch of the theorems' proofs. Detailed proofs can be found in our technical report [14].

II. Related Work

The concept of exploiting FEC for reliable multicasting/broadcasting[1] has been the subject of considerable amount of work, both in wireline and wireless settings. We only survey here work that is most closely related. Rubenstein *et al* [15] propose a multicast protocol that requires a source to forward redundant packets in advance. This protocol is shown to achieve a significant decrease in the expected time for reliable delivery of data. Rizzo *et al* propose RMDP [16], another FEC-based reliable multicast protocol, and show that FEC effectively reduces the amount of acknowledgments. However, the problem of quantifying the level of redundancy remains unsolved.

Huitema [17] and Nonnenmacher *et al* [18] evaluate the performance improvements achieved with different levels of FEC redundancy via numerical computation. Ghaderi *et al* [11] and Mosko *et al* [19] obtain numerical evaluation of the distribution of the completion time. However, no closed form is provided to relate the redundancy needed with the probability of success. Eryilmaz *et al* [20] provide recursive expression for the *average* completion time and Ghaderi *et al* [11] derive an asymptotic expression for it. However, they do not provide results for the CDF. Furthermore, to our knowledge, our work is the first to demonstrate the phase transition associated with this CDF and to derive bounds on the asymptotic error for the case of a finite number of receivers.

While in practice the packet loss probability differs from node to node due to many factors (i.e., link quality, distance to the source, antenna sensitivity) all the following references [3, 8–11, 15–17] assume homogeneous packet loss rates in their analysis. The work in [18, 19] provide analysis for heterogeneous packet loss probability scenarios. However, unlike our paper, the results are only numerical.

III. Background

A. Extreme Value Theory

Let X_1, X_2,.., X_N be independent, identically distributed (i.i.d.) random variables. Extreme value theory provides tools for characterizing possible limit distributions of sample maxima of the above i.i.d. random variables. Denote by F the CDF of X and by F^N the CDF of the maximum of X_1, X_2,.., X_N. Suppose there exists a sequence of constants a_N and b_N, such that $\frac{\max(X_1,X_2,...,X_N)-b_N}{a_N}$ has a nondegenerate limit distribution as $N \to \infty$, then

$$\lim_{N\to\infty} F^N(a_N x + b_N) = G(x), \tag{1}$$

or equivalently,

$$\lim_{N\to\infty} N\bar{F}(a_N x + b_N) = -\log G(x), \tag{2}$$

where $G(x)$ is the CDF of one of the three possible extreme value distributions, namely Fréchet, Gumbel and Weibull [12, p.9].

For a given random variable, various tests exist to determine its *domain of attraction* (i.e., the corresponding extreme value distribution) and its normalization constants. In our paper, all the distributions of interest belong to the domain of attraction of the Gumbel distribution, i.e.,

$$G(x) = \exp(-e^{-x}) \quad \forall x \in \Re. \tag{3}$$

Under mild technical assumptions [12, p.77], the domain of attraction conditions imply also moment convergence. Thus, for distributions belonging to the Gumbel's domain of attraction

$$\lim_{N\to\infty} \frac{\mathbb{E}(\max(X_1, X_2, ..., X_N)) - b_N}{a_N} = \gamma, \tag{4}$$

where $\gamma \approx 0.5772$ is the Euler's constant.

B. Convergence Metric

As mentioned above, an estimation based on EVT assumes $N \to \infty$. We provide now a metric to study the quality of convergence when N is finite. Specifically, we fix a value on the y-axis and measure the distance on the x-axis between the points corresponding to the exact distribution F^N and the limit distribution G. Specifically, as shown in Fig. 1, let

$$\begin{cases} G(x^*) = y, \\ F^N(a_N \widetilde{x} + b_N) = y, \end{cases} \tag{5}$$

then the convergence metric is set as follows

$$\Delta = |x^* - \widetilde{x}|. \tag{6}$$

In the following section, we will derive a bound on Δ that applies uniformly to an entire interval $[y_l, y_h]$, where $0 \le y_l \le y_h < 1$. If the desired completion probability y is known in advance, then the values of y_l and y_h can simply be set to y leading to a tighter bound on Δ. Otherwise, one can select a larger interval and the bound will apply to all values of y belonging to that interval.

[1] The terms multicasting and broadcasting are used interchangeably in this paper.

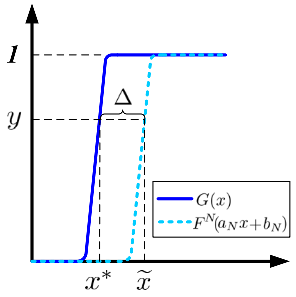


Fig. 1. Illustration of convergence metric.

IV. The Homogeneous Case: Limit Distribution and Convergence Bounds

A. Model and Problem Formulation

We consider the problem of broadcasting a file consisting of M packets from a source (e.g., a base station) to N nodes within its transmission range. The time axis is slotted and each packet transmission takes one time slot. In this section, we assume that each node experiences the same packet loss probability p, independent of any other events.

We assume that FEC is implemented using a perfect rateless code, i.e., each node needs to correctly receive M distinct packets to recover a file. Thus, a source transmits new packets until all the nodes received M different packets. If slightly more packets are needed (say M') because of the imperfection of codes, then one just need to replace M by M' in the following analysis.

Denote by T the random variable representing the completion time, i.e. the number of time slots, needed to disseminate M packets to a cluster of N nodes. Our goal is to characterize the CDF of T, namely $\Pr\{T \le t\}$, with which one can determine the number of redundant packets needed in FEC. Towards this end, we will use EVT to characterize the limiting form of the CDF of T when $N \to \infty$ and then derive bounds on the error Δ for finite values of N.

In this paper, we do not enter into the details of how to estimate the network parameters (i.e., N and p). We refer the interested readers to [4, 21] for possible approaches.

B. Asymptotic Analysis of Completion Time

Denote by T_n^m the number of slots needed for node n to receive its m-th packet, $1 \le m \le M$. Clearly, T_n^m follows a geometric distribution with mean $1/(1-p)$, i.e., $\Pr\{T_n^m = i\} = p^{i-1}(1-p)$.

Thus, the time T_n needed for node n to receive M different packets is the sum of M i.i.d geometric random variables with mean $1/(1-p)$, i.e., $T_n = \sum_{m=1}^{M} T_n^m$ and T_n is said to follow a *negative binomial* or *Pascal* distribution [22, p.166]. Due to the broadcast nature of the channel, the completion time for broadcasting a file to all the nodes is the maximum of N negative binomial random variables, i.e., $T = \max(T_1, T_2, \ldots\ldots T_N)$. The following theorem tightly bounds the distribution of T as $N \to \infty$. Before proceeding, we recall the definition of stochastic ordering [23, p. 404].

Definition 1: A random variable X is stochastically larger than a random variable Y, denoted $X \ge_{st} Y$, if

$$\Pr(X > a) \ge \Pr(Y > a), \qquad \text{for all } a. \tag{7}$$

Theorem 1: The completion time T to disseminate M packets to N nodes using FEC broadcasting is bounded by random variables belonging to Gumbel's domain of attraction. Namely, there exist T_l and $T_u = T_l + 1$, satisfying

$$T_l \le_{st} T \le_{st} T_u,$$

$$\lim_{N\to\infty} \Pr\{(T_l - b_N)/a_N \le x\} = G(x),$$

$$\text{where, } a_N = 1/\log(\frac{1}{p}), \tag{8}$$

$$\begin{aligned} b_N &= \log_{\frac{1}{p}}(N) + (M-1)\log_{\frac{1}{p}}(\tau) \\ &\quad + (M-1)\log_{\frac{1}{p}}(\frac{1-p}{p}) - \log_{\frac{1}{p}}(M-1)!, \end{aligned} \tag{9}$$

$$\tau = \log_{\frac{1}{p}}(N) + (M-1)\left(\log_{\frac{1}{p}}(\frac{1-p}{p})\right). \tag{10}$$

Proof: Since T_n follows a negative binomial distribution, let $D(t) = \Pr\{T_n \le t\} = I(1-p; M, t-M+1)$ [22], where t is an integer and $I(z; a, b)$ is the regularized beta function. Create a continuous R.V. T_u^i with CDF $F(x) = I(1-p; M, x-M+1)$, where $x > M-1$. Let $T_l^i = T_u^i - 1$. The probability distribution function for T_l^i is thus $F(x+1)$. From [24, p. 516, 594, 597],

$$\bar{F}(x) = 1 - I(1-p; M, x-M+1) \tag{11}$$

$$= \frac{\Pi_{j=0}^{M-1}(x-j)}{(M-1)!} p^{x-M+1} \sum_{i=0}^{M-1} \binom{M-1}{i} \frac{(-1)^i p^i}{x-M+1+i}. \tag{12}$$

Since $I(x)$ is an increasing function of x, we have

$$\bar{F}(x+1) \le \bar{D}(\lceil x \rceil) \le \bar{F}(x). \tag{13}$$

Let $T_u = \max_{i=1..N} T_u^i$, and $T_l = \max_{i=1..N} T_l^i = T_u - 1$. By Definition 1, we have $T_l \le_{st} T \le_{st} T_u$.

Inserting Eq. (8) and Eq. (9) into Eq. (12), one can show

$$\lim_{N\to\infty} N\bar{F}(a_N x + b_N) = -\log G(x), \quad x \in R.$$

Thus, F is in the domain of attraction of G with normalizing constants a_N, b_N, namely

$$\lim_{N\to\infty} \Pr\{(T_l - b_N)/a_N \le x\} = G(x).$$

■

Theorem 1 shows that as $N \to \infty$, the CDF of the completion time converges to a scaled and shifted Gumbel distribution. Namely,

$$\Pr\{T \le t\} \approx G(\frac{t - b_N}{a_N}). \tag{14}$$

Since the packet loss probability p is usually small, a_N is pretty small as well and the completion time distribution has a sharp phase transition around the point b_N. This will be verified by our numerical results in Section VI.

A corollary from the theorem is that the performance of rateless coding on a single channel is identical to that of plaintext coding over an unlimited number of channels. As such, Theorem 1 is similar to Theorem 11 in [3]. However, the normalizing constant b_N here is different, which allows for faster convergence.

C. Convergence Bounds

Theorem 1 characterized the limiting form of the CDF of T as $N \to \infty$. The following theorem bounds the asymptotic error using the convergence criterion defined in Eq. (6), that is, it bounds the distance between the asymptotic estimate $\widetilde{x}$ and the exact value x^*.

Theorem 2: The distance $\Delta = |\widetilde{x} - x^*|$ between the exact distribution $\Pr\{T \leq a_N\widetilde{x} + b_N\} = y$ and the Gumbel distribution $G(x^*) = y$ is bounded as follows for all probability values y belonging to the interval $[y_l, y_h]$:

$$\Delta \leq |\Delta_l| + |\Delta_h + \log(1 + \frac{1}{N}\log\frac{1}{y})|, \tag{15}$$

$$\begin{aligned}\text{where } \Delta_l =& \log p - \frac{(M-1)^2}{2(a_N G^{-1}(y_l) + b_N) - M + 3} \\ &+ (M-1)(1 - \frac{\tau}{a_N G^{-1}(y_l) + b_N + 1}),\end{aligned} \tag{16}$$

$$\begin{aligned}\Delta_h =&(M-1)\log(\frac{1}{1-p}) \\ &+ (M-1)\frac{a_N G^{-1}(y_h) + b_N + 1 - \tau}{\tau},\end{aligned} \tag{17}$$

and G^{-1} is the inverse function of G.

Proof: We first provide bounds for $\bar{F}(x)$, defined in Eq. (12). We are going to show

$$\frac{(1-p)^{M-1}}{x-M+1} \leq \sum_{i=0}^{M-1}\binom{M-1}{i}\frac{(-1)^i p^i}{x-M+1+i} \leq \frac{1}{x-M+1}. \tag{18}$$

Since

$$\begin{aligned}&t^{x-M}(1-pt)^{M-1} && (19)\\ &= \sum_{i=0}^{M-1}\binom{M-1}{i}(-1)^i p^i t^{x-M+i} && (20)\\ &= \frac{d}{dt}\left(\sum_{i=0}^{M-1}\binom{M-1}{i}\frac{(-1)^i p^i t^{x-M+1+i}}{x-M+1+i}\right), && (21)\end{aligned}$$

we have

$$\sum_{i=0}^{M-1}\binom{M-1}{i}\frac{(-1)^i p^i}{x-M+1+i} = \int_0^1 t^{x-M}(1-pt)^{M-1}\,dt. \tag{22}$$

For $0 < p < 1$ and $0 \leq t \leq 1$, we have $(1-p)^{M-1} \leq (1-pt)^{M-1} \leq 1$ and Eq. (18) follows.

We next insert Eq. (18) into Eq. (12) to obtain the following bounds on $\bar{F}(x)$

$$\bar{F}(x) \geq \frac{\Pi_{j=0}^{M-2}(x-j)}{(M-1)!}p^{x-M+1}(1-p)^{M-1}, \tag{23}$$

$$\bar{F}(x) \leq \frac{\Pi_{j=0}^{M-2}(x-j)}{(M-1)!}p^{x-M+1}. \tag{24}$$

We now provide bounds on the product $\Pi_{j=0}^{M-2}(x-j)$ of Eq. (23) and Eq. (24), as follows. Let $w = \Pi_{j=0}^{M-2}(x-j)$, we have,

$$\begin{aligned}\log(w) &= \sum_{j=0}^{M-2}\log(x-j) && (25)\\ &\geq \int_0^{(M-2)+1}\log(x-t)\,dt && (26)\\ &= \log\left(\frac{x^x}{(x-M+1)^{x-M+1}}e^{-(M-1)}\right). && (27)\end{aligned}$$

Similarly,

$$\log(w) \leq \log\left(\frac{(x+1)^{x+1}}{(x-M+2)^{x-M+2}}e^{-(M-1)}\right). \tag{28}$$

According to Ref. [25], for $x > 0$, $\log(1+x) \geq \frac{x}{1+\frac{1}{2}x}$. Therefore, one can show

$$\frac{x^x}{(x-M+1)^{x-M+1}} \geq e^{\frac{2(M-1)(x-M+1)}{2x-M+1}}x^{M-1}. \tag{29}$$

According to Ref. [26, p. 242], we have, $\left(1 - \frac{x}{n}\right)^n \leq e^{-x}$. Therefore,

$$\frac{(x+1)^{x+1}}{(x-M+2)^{x-M+2}} \leq e^{M-1}(x+1)^{M-1}. \tag{30}$$

Thus, with Eq. (23), Eq. (27) and Eq. (29), we have

$$\bar{F}(x) \geq \frac{p^{x-M+1}(1-p)^{M-1}}{(M-1)!}e^{-\frac{(M-1)^2}{2x-M+1}}x^{M-1}. \tag{31}$$

Similarly, with Eq. (24), Eq. (28) and Eq. (30), we have,

$$\bar{F}(x) \leq \frac{p^{x-M+1}}{(M-1)!}(x+1)^{M-1}. \tag{32}$$

Now, back to the negative binomial distribution $D(\lceil x \rceil)$. By Eq. (13), Eq. (31) and Eq. (32), one can show

$$\bar{D}(\lceil x \rceil) \geq \frac{p^{x-M+2}(1-p)^{M-1}}{(M-1)!}e^{-\frac{(M-1)^2}{2x-M+3}}(x+1)^{M-1}, \tag{33}$$

$$\bar{D}(\lceil x \rceil) \leq \frac{p^{x-M+1}}{(M-1)!}(x+1)^{M-1}. \tag{34}$$

Note that by Ref. [25], we have,

$$\frac{x}{1+x} \leq \log(1+x) \leq x, \forall x > -1. \tag{35}$$

According to Eq. (33) and Eq. (34) and by the definition of a_N and b_N, one can use Eq. (35) to show

$$\frac{1}{N}e^{-(x-\Delta_l)} \leq \bar{D}(\lceil a_N x + b_N \rceil) \leq \frac{1}{N}e^{-(x-\Delta_h)}. \qquad (36)$$

Since $\Pr\{T \leq a_N x + b_N\} = (D(\lceil a_N x + b_N \rceil))^N$, again by Eq. (35), one can show

$$\exp\left(\frac{-e^{-(x-\Delta_h)}}{1-\frac{e^{-(x-\Delta_h)}}{N}}\right) \leq \Pr\{T \leq a_N x + b_N\} \leq e^{-e^{-(x-\Delta_l)}}. \qquad (37)$$

Since all the three functions are monotonically increasing functions of x, if we let

$$\exp\left(\frac{-e^{-(x_1-\Delta_h)}}{1-\frac{e^{-(x_1-\Delta_h)}}{N}}\right) = \Pr\{T \leq a_N \widetilde{x} + b_N\} = e^{-e^{-(x_2-\Delta_l)}} = y \qquad (38)$$

Then $x_2 \leq \widetilde{x} \leq x_1$. Solving Eq. (38) completes the proof of the theorem. ■

Theorem 2 provides a means to conservatively implement FEC for finite values of N, that is, if one wants to guarantee a completion probability y, then the source should transmit at least $\lceil a_N(x^* + \Delta) + b_N \rceil$ packets. The bound provided by Eq. (15) also exhibits the desirable property of becoming tighter as the number of recipient nodes N increases and as the completion probability y approaches 1.

V. The Heterogeneous Case: Expectation Limit

In this section we relax the assumption of homogeneous packet loss probabilities. Let R be the source's transmission range. Recipient nodes are deployed uniformly at random within a disk of radius of R, with the source at the origin. The signal quality is discretized into L levels based on the distance from the source. Denote by $\overrightarrow{\alpha} = [\alpha_1, \alpha_2, .., \alpha_L]$ the distance vector (normalized by R), where $0 < \alpha_1 < .. < \alpha_L = 1$. Next, let $\overrightarrow{\omega_\alpha} = [\omega_{\alpha_1}, \omega_{\alpha_2}, .., \omega_{\alpha_L}]$ be the corresponding packet loss vector, where $0 < \omega_{\alpha_1} < .. < \omega_{\alpha_L} < 1$. The packet loss probability for a node is ω_{α_l} if its distance from the source is between $\alpha_{l-1}R$ and $\alpha_l R$ ($\alpha_0 = 0$ by definition). This radio model is illustrated in Fig 2. Then, the CDF of the packet loss probability for node n is a multi-step function defined as follows

$$P(p_n \leq x) = \begin{cases} 0 & 0 < x < \omega_{\alpha_1}, \\ \alpha_l^2 & \omega_{\alpha_l} \leq x < \omega_{\alpha_{l+1}},\ l = 1, .., L-1 \\ 1 & x \geq \omega_{\alpha_L}. \end{cases} \qquad (39)$$

Note that the radio model of the previous section is a special case of this model, by setting $L = 1$ and $\omega_{\alpha_L} = p$.

In the remainder of this section, we first establish a relation between the FEC data broadcasting problem in wireless networks and the multi-set coupon collector's problem. This connection enables us in the second part of the section to leverage recent analytical results on the asymptotic behavior of heterogeneous coupon collector systems [13] to analyze FEC data broadcasting with heterogeneous packet loss probabilities.

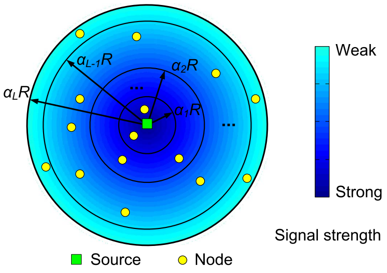


Fig. 2. Radio model for heterogeneous packet loss.

A. Relation between Coupon Collector and Data Broadcasting Problems

In the multi-set coupon collector's problem, a shopper tries to collect M complete sets of N different coupons in several attempts. Coupon n is associated with a value $q_n > 0$. Each attempt provides the collector with a coupon n with probability $q_n / \sum_{i=1}^{N} q_i$. Assume that there is unlimited supply of coupons of each kind. Let $\bar{E}(\{q_n\})$ be the expected number of attempts the collector needs to make in order to obtain M complete sets of N coupons, where $\{q_n\}$ represents the set of coupon's values. Asymptotic limits of this expectation for large values of N are studied in [13].

Back to our original problem, let $\{p_n\}$ be a set of packet loss probabilities associated with each receiver and $\bar{T}(\{p_n\})$ be the expected time to transmit M packets to N users with packet loss probabilities $p_1, p_2, .., p_N$ using FEC data broadcasting. The following Theorem establishes a relation between $\bar{T}(\{p_n\})$ and $\bar{E}(\{q_n\})$.

Theorem 3: For arbitrary packet loss probabilities p_n, $n = 1, .., N$, as $N \to \infty$,

$$\bar{T}(\{p_n\}) = \frac{1}{\sum_{n=1}^{N} \log(1/p_n)} \bar{E}(\{\log(1/p_n)\}) + \Theta(1). \qquad (40)$$

Proof: By embedding the coupon collector's problem into a Poisson process [13], one can show

$$\frac{\bar{E}(\{q_n\})}{\sum_{n=1}^{N} q_n} = \int_0^\infty \left(1 - \Pi_{n=1}^{N}\left(1 - \sum_{i=0}^{M-1} \frac{(q_n t)^i}{i!} e^{-q_n t}\right)\right) dt. \qquad (41)$$

For the data broadcasting problem, by stochastic ordering, one can show, with Y_n being independent Erlang-M random variable with rate equal to $\log(\frac{1}{p_n})$, $n = 1, 2, .., N$,

$$\max_{n=1,..,N} Y_n \leq_{st} T(\{p_n\}) \leq_{st} \max_{n=1,..,N} Y_n + M. \qquad (42)$$

Since Y_n is an Erlang random variable, we have,

$$\bar{F}_{Y_n}(x) = \sum_{i=0}^{M-1} \frac{(\lambda_n x)^i}{i!} e^{-\lambda_n x},\ \lambda_n = \log(1/p_n), \qquad (43)$$

and,

$$\mathbb{E}\max_{n=1,..,N} Y_n$$
$$= \int_0^\infty \left(1 - \Pi_{n=1}^N \left(1 - \sum_{i=0}^{M-1} \frac{(\lambda_n t)^i}{i!} e^{-\lambda_n t}\right)\right) dt. \qquad (44)$$

The Theorem follows from Eq. (41), Eq. (42), and Eq. (44). ■

B. Asymptotic Limit of Expected Completion Time

We next derive a closed-form expression for the expected completion time as $N \to \infty$, for the heterogeneous radio model presented at the beginning of the section.

Theorem 4: If the CDF of the packet loss probability of each node n, $n = 1,..,N$, satisfies Eq. (39), then

$$\lim_{N\to\infty} \frac{\bar{T}(\{p_n\}) - b_N}{a_N} = \gamma, \qquad (45)$$

where $\gamma \approx 0.5772$ is the Euler's constant and

$$a_N = 1/\log(\frac{1}{p'}), \qquad (46)$$
$$b_N = \log_{\frac{1}{p'}}(N') + (M-1)\log_{\frac{1}{p'}}\tau$$
$$+ (M-1)\log_{\frac{1}{p'}}\frac{1-p'}{p'} - \log_{\frac{1}{p'}}(M-1)!, \qquad (47)$$
$$\tau = \log_{\frac{1}{p'}}(N') + (M-1)\left(\log_{\frac{1}{p'}}\frac{1-p'}{p'}\right), \qquad (48)$$
$$p' = \omega_{\alpha_L}, \quad N' = (\alpha_L^2 - \alpha_{L-1}^2)N. \qquad (49)$$

Proof: We first compute the expected time to obtain all coupons, $\bar{E}(\{\log(1/p_n)\})$. According to Ref. [13], with $Y_1, Y_2, .., Y_N$ being i.i.d. random variables having Erlang-M distribution with rate parameter equal to 1, we have

$$P\left(\frac{Y_n}{q_n} > s\right) = \int_0^\infty P(q_n \le \frac{x}{s})\frac{x^{M-1}e^{-x}}{(M-1)!}dx. \qquad (50)$$

With $q_n = \log\frac{1}{p_n}$, $n = 1,..,N$ and the CDF of p_n described in Eq. (39), one can compute the CDF of q_n and show

$$P\left(\frac{Y_n}{q_n} > s\right) = \sum_{l=1}^{L}\sum_{i=0}^{M-1}(\alpha_l^2 - \alpha_{l-1}^2)\frac{(s\log\frac{1}{\omega_{\alpha_l}})^i}{i!}e^{-s\log\frac{1}{\omega_{\alpha_l}}}. \qquad (51)$$

Let $F_c(s) = 1 - P\left(\frac{Y_n}{q_n} > s\right)$, $F_c'(s) = \frac{d}{ds}F_c(s)$, and set a_N and b_N as in Eq. (46) and Eq. (47). It can be shown as $N \to \infty$

$$\begin{cases} \frac{1}{N} = 1 - F_c(b_N), \\ a_N = \frac{1-F_c(b_N)}{F_c'(b_N)}. \end{cases} \qquad (52)$$

According to Theorem 4.1 of Ref. [13], if Eq. (52) holds, then

$$\lim_{N\to\infty} P\left(\frac{\frac{\bar{E}(\{q_n\})}{N\mu} - b_N}{a_N} \le x\right) = G(x), \qquad (53)$$
$$\lim_{N\to\infty}\frac{\frac{1}{N\mu}\bar{E}(\{q_n\})) - b_N}{a_N} = \gamma, \qquad (54)$$

where μ is the expected value of random variable q_n.

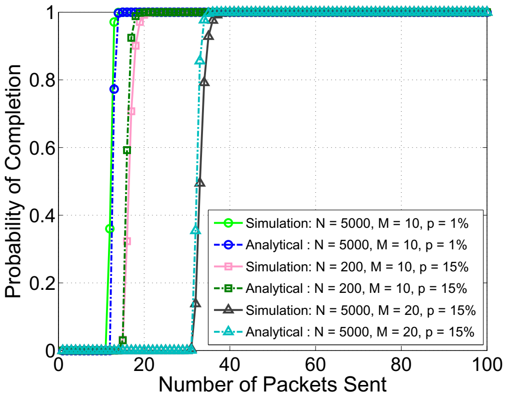


Fig. 3. Accuracy of asymptotic estimate and phase transition demonstration.

By Theorem 3, we have,

$$\lim_{N\to\infty}\frac{\bar{T}(\{p_n\}) - b_N}{a_N} \qquad (55)$$
$$= \lim_{N\to\infty}\frac{\frac{1}{N\mu}\bar{E}(\{\log(1/p_n)\}) - b_N}{a_N} = \gamma, \qquad (56)$$

and the theorem follows. ■

This theorem provides the following insight. On average the number of nodes staying within the furthest ring in the disk is $(\alpha_L^2 - \alpha_{L-1}^2)N$. By Eq. (45), $\bar{T}(\{p_i\})$ is asymptotically identical to the expected time needed to disseminate M packets to $(\alpha_L^2 - \alpha_{L-1}^2)N$ nodes with homogeneous packet loss probability ω_{α_L}. Hence, as $N \to \infty$, the time needed to disseminate packets to the nodes with the highest packet loss probability dominates.

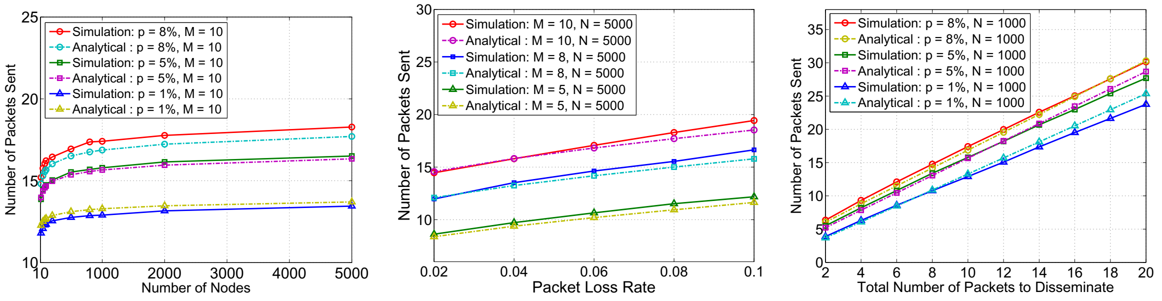
VI. Numerical Results

In this section, we illustrate the major analytical findings of this paper, namely, (i) the accuracy of the asymptotic estimate of the CDF of the completion time provided by Theorem 1, (ii) the phase transition behavior of this CDF, (iii) the tightness of the upper and lower bounds derived along the proof of Theorem 2, and (iv) the accuracy of the asymptotic limit on the expected completed time provided by Theorem 4. All the simulation plots are obtained by averaging results over 10000 simulations with identical parameters, but different random seed.

A. Accuracy of Asymptotic Estimate

Fig. 3 compares the CDF estimated by Theorem 1, with the CDF obtained from simulation for various parameters M, N, and p. It is evident from the figure that the limit form provides an accurate estimate of the actual distribution. It is interesting to note that even with a large number of receivers and relatively high loss packet probability, we do not need a large number of redundant packets to ensure file reception (with high probability) by all the nodes.

Fig. 3 also clearly demonstrates the phase transition behavior of the CDF. As expected, the CDF shifts to the right as the number of nodes N or the number of file packets M increases,

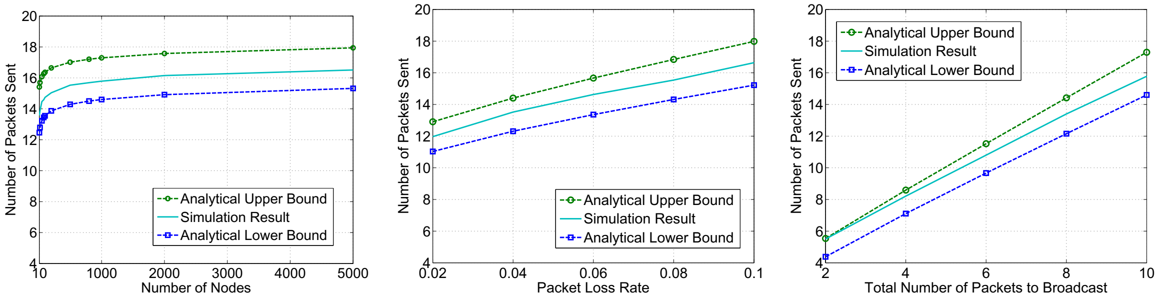


(a) Varying the number of nodes N.

(b) Varying the packet loss rate p.

(c) Varying the number of file packets M.

Fig. 4. Number of packets needed to be sent to guarantee completion with probability 99%: Varying different parameters.



(a) $M = 10$, $p = 5\%$, Varying N.

(b) $N = 5000$, $M = 8$, Varying p.

(c) $N = 1000$, $p = 5\%$, Varying M.

Fig. 5. Number of packets needed to guarantee completion with probability 99%, comparison of simulation and analytical bounds.

but the sharpness phase transition is not much affected. On the other hand, the packet loss rate p has an effect on both translating and scaling the CDF. A smaller value of p shifts the CDF to the left and also results in a sharper transition.

As discussed in Section IV-B, the phase transition occurs around the point b_N. Using Eq. (9), one can compute the values of b_N for the three cases shown in Fig. 3, which are found to be 12.709, 15.659, 32.020. These values accurately locate the phase transition points.

According to Theorem 1, as $N \to \infty$, the CDF converges to a Gumbel distribution scaled by $\frac{1}{a_N}$ and translated to the right by $\frac{b_N}{a_N}$. To verify this finding, we closely examine each parameter by fixing the other two. Results are shown in Fig. 4(a), 4(b) and 4(c). We study the phase transition shift of the CDFs as the parameters change by evaluating the case where the completion probability is 99%.

Fig. 4(a) shows the number of packets needed as N increases. As predicted, when M and p is fixed, the number of packets needed increases logarithmically with N. This is true even when the number of nodes is small, e.g. $N = 10, 20, 50$. This result explains why the redundancy needed is relatively small, even for large values of N.

Fig. 4(b) demonstrates the case where N and M are fixed. According to Theorem 1, given N and M, the number of packets needed is a linear function of $1/\log\frac{1}{p}$. Using Taylor's expansion, for a small value of p, $1/\log\frac{1}{p} \approx 1/(1-p) \approx 1+p$. Thus, the number of packets is approximately a linear function of p, which coincides with Fig. 4(b).

Fig. 4(c) shows that, fixing N and p, the phase transition shifts to the right linearly as M increases. This is because the CDF shifts by $\frac{b_N}{a_N}$, which is close to a linear function of M.

	$\overrightarrow{\alpha}$	$\overrightarrow{\omega_\alpha}$	**N**	**M**
Case 1	**[0.5, 1]**	**[0.1, 0.11]**	**50**	**20**
Case 2	**[0.5, 1]**	**[0.1, 0.15]**	**50**	**20**
Case 3	**[0.9, 1]**	**[0.0498, 0.1353]**	**50**	**20**
Case 4	**[0.2, 0.4, 0.6, 0.8, 1]**	**[0.001, 0.008, 0.027, 0.064, 0.125]**	**50**	**20**
Case 5	**[0.2, 0.4, 0.6, 0.8, 1]**	**[0.001, 0.008, 0.027, 0.064, 0.125]**	**1000**	**20**

Fig. 6. Different network settings for heterogeneous packet loss simulation.

B. Tightness of Bounds

We next compare the analytical upper and lower bounds with simulation results. Each parameter (i.e., N, M, p) is investigated by fixing the other two. Simulation results are compared with analytical bounds, shown in Fig 5(a), Fig 5(b) and Fig 5(c). The analytical bounds are obtained by Eq. (37). We fix $y_l = y_h = 99\%$, that is, a 99% probability of completion. As expected, the curve representing simulation result lies between the analytical lower bound and upper bound. More importantly, the gap between the upper bound and the simulation result is reasonably small for a variety of different parameters. If one is to use the analytical upper bound to estimate the amount of redundant packets, then only one or two more packets than necessary would be transmitted.

C. Heterogeneous Packet Loss

In this part we verify the result of Theorem 4, which states that as $N \to \infty$, the time needed to disseminate packets to the

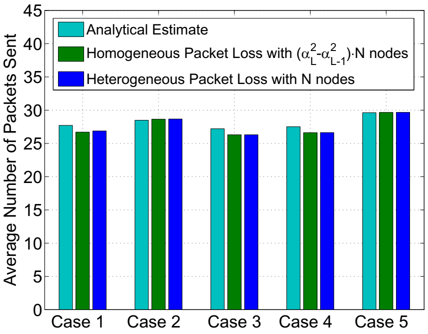


Fig. 7. Average completion time for scenarios with heterogeneous packet loss vs. homogeneous packet loss.

nodes with the highest packet loss probability dominates. We investigate different network settings as shown in Fig. 6.

We compare three results: (i) average time to disseminate M packets to N nodes under heterogeneous packet loss described by distance vector $\overrightarrow{\alpha}$ and packet loss vector $\overrightarrow{\omega_\alpha}$, based on simulation; (ii) the average time to disseminate M packets to $(\alpha_L^2 - \alpha_{L-1}^2)N$ nodes all having the same packet loss probability ω_{α_L}, based on simulation; and (iii) the analytical estimate based on Theorem 4.

The results are shown in Fig. 7. In cases 1 through 3 we assume, $L = 2$ levels of signal quality: nodes within distance $\alpha_1 R$ of the source have lower packet loss probability ω_{α_1}. Nodes beyond this radius have higher packet loss probability ω_{α_2}. The results show that the presence of nodes with lower packet loss rates has little impact on the completion time for the entire network. This is true even when only a small fraction of nodes suffers from higher packet loss rates. For example, in Case 3, on average only $19\% \cdot N$ of nodes have higher packet loss probability e^{-2}, and yet distributing a file only to these nodes takes as much as the time to distribute a file to a network where $81\% \cdot N$ nodes have packet loss probability e^{-3} and $19\% \cdot N$ nodes have packet loss probability e^{-2}. In cases 4 and 5, the signal quality is discretized into $L = 5$ levels and reveal similar behavior. In all cases, Theorem 4 predicts well the simulation results.

VII. Prototype Implementation

A. Set-up

In this section, we describe practical implementation of extreme value FEC into the Rateless Deluge over-the-air programming protocol [4]. This protocol uses random linear codes for data encoding and enables efficient distribution of a new file program to all the nodes of a sensor network.

The default setting of Rateless Deluge is as follows. A file is divided into pages and each page consists of 20 packets, where each packet contains 23 bytes of data. A sensor sends out a request if it discovers its neighbors have new data. The request message specifies the page number and the number of packets it needs. When a sensor receives enough number of packets (in our case 20), it can decode the page successfully. As in the original Deluge protocol [1], a sensor suppresses its request if it overhears similar requests sent recently.

Here, we augment the original Rateless Deluge with extreme value FEC, and refer to the new protocol as Extreme Value FEC Deluge. Extreme Value FEC Deluge operates the same as Rateless Deluge except that when receiving a request for a new page, the base station broadcasts a redundant amount of packets. The redundancy is set to guarantee with high enough probability that all the receivers recovered the file. In our case, we set the desired completion probability to be 97%, and the redundancy is then computed using Theorem 1.

The performance of Rateless Deluge and Extreme Value FEC Deluge are evaluated on a testbed consisting of 20 Tmote Sky sensors (see Fig. 8). All the sensors are within communication range. Sensors transmit at their highest power setting over short distances to ensure a good link, and packet loss at the receiver is forced by dropping packets uniformly at random. One sensor serves as the base station and 18 others are receivers. The last sensor is used to record network traffic. During each experiment, a new file is injected from a PC into the base station and the base station then disseminates it to the network.

B. Results

In our first experiment, we disseminate a single page, 20-packet file using Rateless Deluge. The packet loss probability is $p = 18\%$. We record the number of data packets sent until every node finishes receiving the file. Based on 200 identical iterations, we plot in Fig. 9 the CDF of the number of packets sent and compare it with the analytical estimate from Theorem 1. We observe that the theory predicts well the experimental results. Further, even though the number of sensors in the network is relatively small, the sharp phase transition is still evident.

Next, we compare the performance of Rateless Deluge and Extreme Value FEC Deluge. We distribute a 20-packet file and take averages over 200 identical experiments. We analyze the network traffic in control plane as well as data plane, namely, we record the number of request messages and data messages sent. We also record the completion time to disseminate the file. The results of the comparison are summarized in Fig. 10.

The results show that Extreme Value FEC Deluge sends out slightly more data messages (less than 5%). However, it drastically reduces the amount of request messages by a factor of about five compared to Rateless Deluge. Note that the minimum possible number of request messages is one since at least one request message must be sent to initiate the dissemination process. With Extreme Value FEC Deluge, the average number of requests is 1.225. Thus, most of the time the entire network finishes receiving enough packets after the base station's first set of transmissions. Thanks to its lower control plane overhead, Extreme Value FEC Deluge effectively reduces the completion time to disseminate a 20-packet file to a 18-node network to 1.14 sec, which is about half of the time needed by Rateless Deluge. We observed similar results when disseminating larger files.

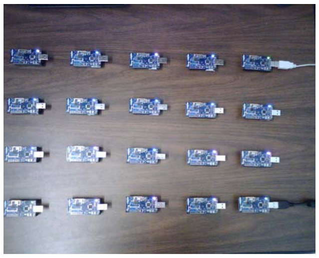


Fig. 8. Experimental testbed with 20 Tmotes.

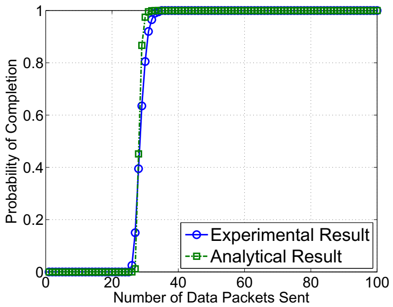


Fig. 9. Real sensor experiments vs. analysis: $N = 18$, $M = 20$, $p = 18\%$.

VIII. Concluding Remarks

In this paper, we have developed theoretical foundations and demonstrated practical use of a highly efficient strategy for reliable data broadcasting, called extreme value FEC. This strategy accurately predicts the number of redundant packets to be disseminated by a source so to avoid (with high probability) unnecessary retransmission requests by receivers.

Our analysis, based on extreme value theory, accurately captures characteristics of the completion time of FEC data broadcasting. Not only does it demonstrate the phase transition of the CDF of the completion time, but also accurately pinpoints the location of the phase transition point. The analysis also reveals that the number of redundant packets required to guarantee file completion by all receivers increases only logarithmically with N. Another major contribution of the paper is in providing convergence bounds for finite N, demonstrating fast convergence of the asymptotic estimate.

By establishing a relation with the multi-set coupon collector's problem, we also provide a closed-form expression on the expected completion time to disseminate a file to receivers with heterogeneous packet loss probabilities. The result points out that, as N gets large, the time needed to disseminate packets to the nodes with the highest packet loss probability dominates. Simulations confirm this finding even when only a small fraction of nodes suffers from high packet loss rates.

Finally, the paper reports a practical implementation of the extreme value FEC strategy in conjunction with the Rateless Deluge OAP protocol. The results show significant performance improvement with respect to control-plane overhead and average data dissemination time, thereby validating the benefits of our approach under real network settings.

	Rateless Deluge	EV-FEC Deluge
Average Number of Requests	5.685	1.225
Average Number of Data Packets	29.120	30.270
Completion Time (sec)	2.215	1.140

Fig. 10. Rateless Deluge vs. EV-FEC Deluge: 1 page, $N = 18$, $M = 20$, $p = 18\%$.

References

[1] J. Hui and D. Culler, "The dynamic behavior of a data dissemination protocol for network programming at scale." in *SenSys'04*, Nov. 2004.

[2] W. Xiao and D. Starobinski, "Poster abstract: Exploiting multi-channel diversity to speed up over-the-air programming of wireless sensor networks," in *SenSys'05*, San Diego, California, USA, Nov. 2005.

[3] D. Starobinski, W. Xiao, X. Qin, and A. Trachtenberg, "Near-optimal data dissemination policies for multi-channel, single radio wireless sensor networks," in *IEEE INFOCOM 2007*, Anchorage, May 2007.

[4] A. Hagedorn, D. Starobinski, and A. Trachtenberg, "Rateless deluge: Over-the-air programming of wireless sensor networks using random linear codes," in *IPSN 2008*, Saint Louis, MO, USA, Apr. 2008.

[5] Y.-C. Tseng, S.-Y. Ni, Y.-S. Chen, and J.-P. Sheu, "The broadcast storm problem in a mobile ad hoc network," in *Wireless Networks*, vol. 8, no. 2/3. Kluwer Academic Publishers, 2002, pp. 153–167.

[6] P. Levis, N. Patel, S. Shenker, and D. Culler, "Trickle: A self-regulating algorithm for code propagation and maintenance in wireless sensor networks," University of California at Berkeley, Tech. Rep., 2004.

[7] N. Shacham and P. McKenney, "Packet recovery in high-speed networks using coding and buffer management," *INFOCOM '90*, Jun 1990.

[8] Y. Bartal, J. Byers, M. Luby, and D. Raz, "Feedback-free multicast prefix protocols," *ISCC '98.*, pp. 135–141, 1998.

[9] J. Byers, M. Luby, and M. Mitzenmacher, "A digital fountain approach to asynchronous reliable multicast," *Selected Areas in Communications, IEEE Journal on*, vol. 20, no. 8, pp. 1528–1540, Oct 2002.

[10] M. Ghaderi, D. Towsley, and J. Kurose, "Network coding performance for reliable multicast," *MILCOM 2007. IEEE*, pp. 1–7, Oct. 2007.

[11] ——, "Reliability gain of network coding in lossy wireless networks," *INFOCOM 2008.*, April 2008.

[12] S. I. Resnick, *Extreme Values, Regular Variation, and Point Processes*. Springer, 1987.

[13] L.Holst, "Extreme value distributions for random coupon collector and birthday problems," *Extremes*, vol. 4, no. 2, pp. 129–145, 2001.

[14] W. Xiao and D. Starobinski, "Extreme value FEC for wireless data broadcasting," in *Center for Information and Systems Engineering Technical Report: 2008-IR-0050*, Boston University, 2008.

[15] D. Rubenstein, J. Kurose, and D. Towsley, "Real-time reliable multicast using proactive forward error correction," in *NOSSDAV'98*, 1998.

[16] L. Rizzo and L. Vicisano, "RMDP: an FEC-based reliable multicast protocol for wireless environments," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 2, no. 2, pp. 23–31, 1998.

[17] C. Huitema, "The case for packet level FEC," in *Protocols for High-Speed Networks'96*. Chapman & Hall, Ltd., 1996, pp. 109–120.

[18] J. Nonnenmacher, E. Biersack, and D. Towsley, "Parity-based loss recovery for reliable multicast transmission," in *SIGCOMM '97*, 1997.

[19] M. Mosko and J. J. Garcia-Luna-Aceves, "An analysis of packet loss correlation in FEC-enhanced multicast trees," in *ICNP '00*, 2000.

[20] A. Eryilmaz, A. Ozdaglar, and M. Medard, "On delay performance gains from network coding," *CISS*, 2006.

[21] D. S. J. De Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," in *MobiCom '03*, 2003, pp. 134–146.

[22] W. Feller, *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, Inc., 1968, vol. 1.

[23] S. Ross, *Stochastic Processes*, 1996.

[24] M. Fogiel and J. R. Ogden, *Handbook of Mathematical, Scientific, and Engineering Formulas, Tables, Functions, Graphs, Transforms*. Research & Education Assoc., 1984.

[25] E. R. Love, "Some logarithm inequalities," *The Mathematical Gazette*, vol. 64, no. 427, pp. 55–57, 1980.

[26] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*. Cambridge University Press, 1979.