Quality-Aware Routing in Time-Varying Wireless Networks

C. Emre Koksal and Hari Balakrishnan[*]

Abstract

This paper considers the problem of selecting good paths in an ad hoc wireless network. It is well-known that picking the shortest path, in terms of number of hops, between two nodes often leads to poor performance, because such paths tend to use a smaller number of long-range links that could have marginal quality. As a result, *quality-aware* routing metrics are desired for networks that built solely from wireless radios. Previous work, such as De Couto et al.'s ETX, has developed metrics that work well when wireless channel conditions are relatively static, but typical wireless channels experience variations at many time-scales. For example, channels may have low average packet loss ratios, but with high variance, implying that metrics that use the mean loss ratio will perform poorly. In this paper, we describe a new metric, called ENT (Effective Number of Transmissions) that works well under a wide variety of channel conditions. In addition to analyzing and evaluating the performance of ENT, we provide a unified geometic interpretation for wireless quality-aware routing metrics.

1. Introduction

This paper considers the problem of selecting good paths in a mobile ad hoc wireless or sensor network. It is well-known that picking the shortest path, in terms of number of hops, between two nodes often leads to poor performance, because such paths tend to use a smaller number of long-range links that could have marginal quality. As a result, *quality-aware routing (QAR)* metrics are desired for networks that are built using wireless radios.

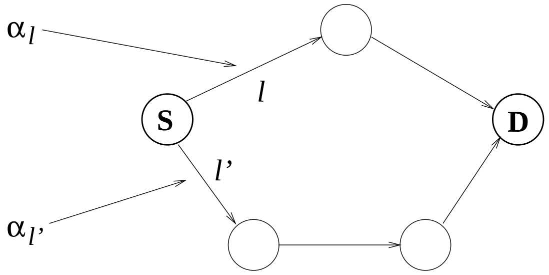

Figure 1: Steps of routing problems include assigning a cost to each link and finding the best link.

In wired networks, the routing problem can be modeled as a graph where nodes are connected by edges of certain weights, on which solving a network optimization problem (*e.g.,* shortest-paths) gives the paths to use to send data between nodes. However, in a network with wireless links, we cannot talk about "links" in the same sense as in wired networks. Wireless communication using radio is time-varying and radio range is unpredictable, and depends greatly on the other radio communications occurring concurrently elsewhere in the network. Moreover, radio communication quality depends on noise and channel fading, especially when nodes move. All of these factors cause high variability in the quality of the "link" between any two nodes, and makes it hard to adapt traditional wired routing techniques to wireless networks.

Finding the best paths between nodes in a wireless network involves three steps:

1. *Assigning metrics to links.* Due to the large size of the parameters that affect a link, this task is non-trivial. Moreover, it is desirable that these metrics are *composable* so that the end-to-end metric of a path can be easily derived from the metrics of the links on the path.
2. *Determining the best path from the link metrics.* The optimum path can be the one that minimizes the total cost or the one with the minimum $\max_{l \in \text{link}} \alpha_l$ where α_l is the cost of forwarding a packet over link l as shown in Fig. 1.
3. *Disseminating routing information.* The link and/or path metrics need to be disseminated in some manner to help nodes select paths by evaluating path metrics. Several routing protocols for mobile networks have been developed in the past (*e.g.,* [1, 2], etc.) to cope with mobility, and changing topologies.

This paper presents a detailed investigation of the first step,

[*] The authors are with the MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA. E-mail:{emre,hari}@csail.mit.edu

determining link metrics, by developing a quality-aware routing metric that captures both long-term link quality and short-term variability of the radio channel between nodes. Our metric is called **ENT** (*Effective Number of Transmissions*), and represents the cost (in terms of number of the number of transmissions) of sending a packet over a link in such a way that the path that minimizes the overall cost while providing guarantees on the maximum packet loss probability visible to higher layers (*e.g.,* TCP connections) that use the path. Our metric takes into consideration not only the physical layer parameters, but also the application level loss rate requirements. We emphasize that, like many other cost metrics, the ENT does not require a change in the structure of the underlying optimization problem that constructs the optimum paths.

The rest of the paper is organized as follows. In the next section, we give motivation for the work and compare our work to previous QAR metrics such as ETX [3]. In Section 3, we present our network model, develop the ENT metric and give a geometric interpretation of different QAR metrics to show how they relate to one another. An important requirement for any QAR metric is that the node be able to estimate channel conditions properly; we discuss the key ideas underlying good channel estimation in Section 4. The paper concludes with a summary of our contributions Section 5.

2. Motivation and Related Work

The traditional approach to routing in ad-hoc wireless networks has been minimum-hop routing [4, 1]. The simplicity of minimum-hop routing is attractive in the face of node mobility. However, minimum-hop routing inherently "quantizes" the state of a link into one of the two states, "up" or "down." Several researchers have described why shortest-path routing in wireless networks leads to sub-optimal performance [3, 5]: such routing leads to paths that use longer-range links of marginal quality. To counter these performance problems, researchers have proposed QAR metrics that consider the performance characteristics of the individual radio "links" along a path in deciding how to route packets.

We observe that the type of the QAR metric to be chosen depends on the physical layer being used. First, suppose that each node uses robust error control coding along with power control to achieve low error rates and adapts to changing channel conditions. In this scenario, low packet loss ratios can be achieved throughout the network. In this case, a uniformity in the link quality and reliable transmission rates is possible, which makes minimum-hop routing a reasonable approach. Thus, if the physical layer is able to estimate and adequately adapt its coding scheme to cope with channel variability, QAR is not critical.

Adaptive physical layer schemes to "hide" the vagaries of the wireless links are hard to achieve in practice. In fact, we know of no current or next-generation radios that propose to employ sophisticated techniques to fully handle channel quality issues at the physical layer, because of implementation complexity and the absence of practically useful codes that can perform well (especially in the non-asymptotic limit of finite packet sizes) across the large range of channel conditions that are observed in practice. In addition, if energy consumption were an issue and power-controlled radios were used, shortest-path routing is worse than minimum-energy routing based on estimating channel quality [6].

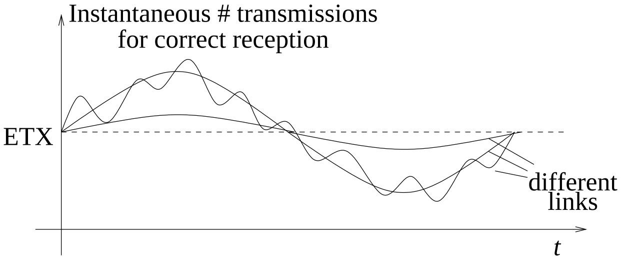

Figure 2: All three links have the same ETX.

Today's practical wireless radios such as the ones based on the 802.11 standard employ only a simple coding strategy, mostly for error detection. Nodes transmit at constant power level and rely on a small number of link-layer packet retransmissions to overcome errors. Some 802.11 systems vary the modulation (and hence the rate) based on current error conditions, slowing the transmission rate when the error rate is high. Our work focuses on QAR over radio networks comprised of radios similar to 802.11 or Bluetooth, where packet losses are visible to higher layers. We use observations of bit-errors and link-layer frame losses to develop a QAR metric.

2.1 Related Work

Our metric is inspired by ETX, a new metric proposed by De Couto et al. [3]. In ETX, each node estimates the packet loss rate p_f to each of its neighbors over a recent time window, and obtains an estimate p_r of the reverse direction from its neighbor (these loss estimates are obtained using broadcast packets that are not retransmitted at the link layer). The node then estimates the *expected transmission (ETX) count* to a neighbor as $\frac{1}{(1-p_f)(1-p_r)}$, and picks the path that has the smallest ETX value from a set of choices.

Yarvis et al. [5] propose a QAR metric that estimates the per-link delivery ratios and uses the product of these as the path metric. This metric does not account for the total bandwidth consumed, because it will prefer two links of low loss rates preferentially over a single link with higher loss-rate; when link-layer retransmissions are used, the single higher-loss link may be able to deliver the packet without as many total transmissions as the two-hop path (ETX is motivated by this observation). Adya et al. [7] propose a delay based QAR metric . This metric uses the measured average round trip time seen by unicast probes between neighboring nodes.

2.2 Channel Variability

Although these QAR metrics, especially ETX, show impressive gains over traditional shortest path routing, they only take the average link behavior into consideration. For instance, consider the three fictional radio links shown in Fig-

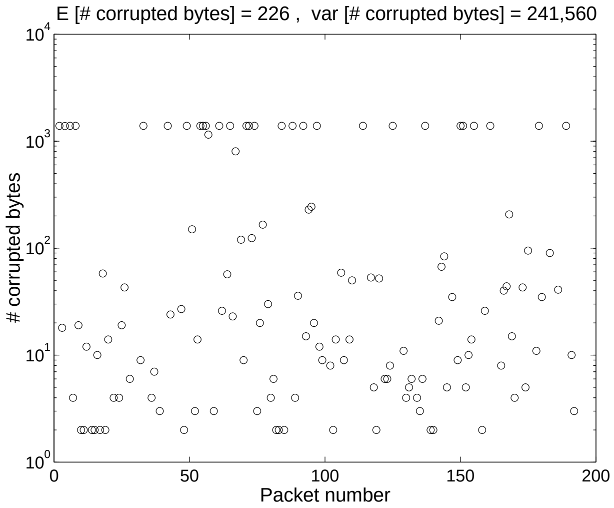

Figure 3: E [# corrupted bytes/packet] = 226, std. dev. # corrupted bytes/packet = 492.

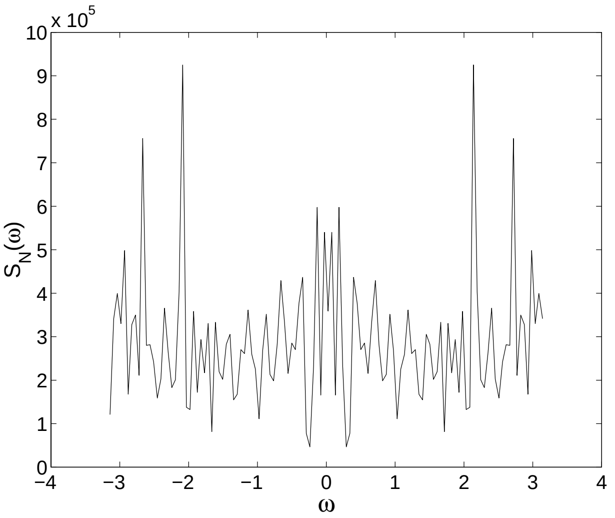

Figure 4: The estimated power spectral density of the trace shown in Figure 3.

ure 2, all of which have the same ETX metric. These links behave in very different ways in time. This example raises two questions:

1. In reality, do we observe such variable behavior in wireless links?[1]
2. If so, to what extent does channel variability affect packet loss behavior?

Several previous measurements have shown that channels vary over short time-scales. Willig et al. present the analysis of 802.11b-based packet traces collected in an industrial environment, showing that packet losses occur in bursts [8]. The coefficient of variation for the number of corrupted consecutive before a successful reception in their trace is about 17, which suggests highly bursty behavior.[2] Woo et al. [9], and Zhao and Govindan [10] have both observed a significant variability in link quality in wireless sensor networks. The former paper points out that the instantaneous packet error probability varies by approximately 30% around its mean. The latter paper, as well as Willig et al., both show that the packet-error stochastic process has long-term dependencies.

There are also a large number of past studies on modeling the instantaneous bit error probability. Most of them use Markov models for this purpose. For instance, Gilbert and Elliot [11, 12] used a two stage Markov chain, Fritchman [13] used a multi-stage Markov chain and Willig [14, 8] used semi-Markov processes and bi-partite models.

We motivate our work by presenting additional anecdotal evidence for channel variability over short time-scales across 802.11b links. This evidence is by no means comprehensive, but is intended to provide some intuition.

The packet trace shown in Figure 3 is taken from a transmission between two nodes, connected using a 1 Mbit/s 802.11b radio, placed 0.55 miles from each other across an open area. The transmitting node sends a continuous stream of 1392-byte ping packets, where each packet has a well-defined byte-pattern. Each point in the graph shows the number of corrupted bytes in each packet. In this trace, the average number of corrupted bytes per packet is 226, with a standard deviation of 492. The link is clearly a marginal one, but also shows how variable link quality can be.

Figure 4 depicts the power spectrum of the number of corrupted bytes in this trace (# corrupted bytes); this graph shows that there is significant power in non-DC frequencies and that certain time scales are much stronger than others. This strongly supports the observation that a wireless link can be highly variable even over short time-scales.

In this paper, we develop the ENT metric which takes not only the average but also the variability of the instantaneous link quality. ENT is a two dimensional metric in the sense that it combines physical channel parameters and application level loss requirements as a single cost.

More precisely, each node estimates the mean and the variance of the bit error probability periodically and combines them with the desired probability of loss rate for the application to evaluate the effective number of transmissions to all of its neighbors. These values are then used by various routing algorithms to find the best path to send packets to a given destination node.

3. Quality-Aware Routing with ENT

3.1 Model

[1] Note that the time-scale over which path-selection decisions are made is typically hundreds of packets; *i.e.,* once a path between two nodes has been selected, it is unlikely to be changed from packet to packet. When we talk about variability, we mean variability over a time-scale of single packet times up to the time to transmit a few hundred packets.

[2] I.e., the ratio of the variance of the number of retransmissions to the square of the expected number of retransmissions is 17.

For the channel, we assume time varying binary symmetric channel. Namely, a bit transmitted at time t is misdetected by the intended receiver with probability $p_{B,t}$. We will not attempt to model $\{p_{B,t},\ t \geqslant 0\}$. We only assume that it is stationary for the time being and will relax this assumption later.

We assume fixed sized packets and let S be the packet size. At a given link, for a packet to be received error free by the receiving node, all of its bits need to be transferred without an error. If an error is detected, packet is dropped and gets retransmitted over the link. Let $p_c(T)$ be the probability that bits $T-S+1, \ldots, T$ are all received correctly. Conditional on $p_{B,t},\ T-S+1 \leqslant t \leqslant T$, $p_c(T)$ can be written as,

$$p_c(T) = \prod_{t=T-S+1}^{T} (1 - p_{B,t}) \tag{1}$$

We define $1/p_c(T)$ as the instantaneous number of transmissions at time T. It physically signifies the average number of transmissions for correct reception if the instantaneous number of transmissions remained fixed at $p_c(T)$. Note that[3]

$$\text{ETX} = \text{E}\left[\frac{1}{p_c(T)}\right] \tag{2}$$

Before the analysis, we discuss some of ETX's shortcomings. In Section 2.2, we discussed how wireless channels vary over both short (single-packet) and longer time-scales. Thus, $\text{E}\left[\frac{1}{p_c(T)}\right]$ is not a good representative of the quality of a wireless channel, because the channel quality is extremely variable, and good or bad states of the channel can be highly persistent. For example, a link with a lower ETX metric may in fact lead to a *higher* observed loss rate at the transport layer, because good link-layer protocols do not try to retransmit lost packets forever but give up after M attempts. When losses occur in bursts, picking a link in the middle of a burst-error situation would be bad *even* if it had a lower ETX. If the goal is to reduce the observed loss rate at the sender's transport layer (*e.g.,* a TCP sender), then ETX may pick a sub-optimal link. The bad choice is likely to be problematic when a node in the network is faced with a choice of two links, one that has a lower ETX but higher variance in $1/p_c(T)$ and the other than has a higher ETX but lower variance in $1/p_c(T)$. Our ENT metric therefore considers both the mean (ETX) and the variance of $1/p_c(T)$.

3.2 The ENT Metric

Instead of comparing the expected number of transmissions for links, we develop a metric for each link that takes into account the probability that the number of transmissions exceed a certain threshold. Our goal is to attempt to reduce the packet loss ratio observed by higher-layer protocols, *after* any link-layer retransmissions are done. The ENT metric involves both the mean and the variance of $1/p_c(T)$, the instantaneous number of transmissions.

First, using Eq. (1) we can write,

$$\begin{aligned} \log \frac{1}{p_c(T)} &= \sum_{t=T-S+1}^{T} \log \frac{1}{1-p_{B,t}} \\ &\approx \sum_{t=T-S+1}^{T} p_{B,t} \end{aligned} \tag{3}$$

where Equation (3) follows when $p_{B,t}$ (the individual bit-error rate) is reasonably small.

Now, consider the probability that the instantaneous number of *link-layer* transmissions exceeds M. Let

$$\begin{aligned} \mu &= \exp\left(\text{E}\left[\log 1/p_c(T)\right]\right) \\ &= \exp\left(S\ \bar{p}_B\right) \end{aligned}$$

where $\bar{p}_B = \text{E}\left[p_{B,t}\right]$ and μ can be thought of as a modification of expected number of transmissions, $\text{E}\left[1/p_c(T)\right]$ (in fact $\mu \leqslant \text{E}\left[1/p_c(T)\right]$ from Jensen's inequality). Also, let

$$\sigma_\Sigma^2 = \text{var}\left(\log\left(1/p_c(T)\right)\right)$$

Due to stationarity, parameters μ and σ_Σ^2 are constant. Finally, let $\rho = \log(M/\mu)$.

The event that the instantaneous number of transmissions exceeds M can be viewed as a threshold crossing for the process $\{p_{B,t},\ t \geqslant 0\}$. Thus, the probability of this event occuring can be written as the probability that the sum of S steps of the process $\{p_{B,t},\ t \geqslant 0\}$ exceeds $\log M$:

$$\begin{aligned} \text{P}\left(\frac{1}{p_c(T)} \geqslant M\right) &= \text{P}\left(\sum_{t=T-S+1}^{T} p_{B,t} \geqslant \log M\right) \\ &\approx \exp\left(-\frac{1}{2}\frac{\rho}{\sigma_\Sigma^2}\cdot\rho\right) \end{aligned} \tag{4}$$

The derivation is based on a large deviation analysis, and is detailed in the appendix. The assumptions we make in the derivation of this result are that $S \gg 1$ is large and $\rho/\sigma_\Sigma^2 \ll 1$. These assumptions are reasonable in practice, as discussed in Section 2.2. The right side of Equation (4) is an upper bound which gets exponentially tight as S increases.

Next, we will incorporate the application level loss rate requirement into the picture. Suppose, it is desried by the application that the probability that the instantaneous number of transmissions exceeds M to be $P_{\text{app}} \leqslant \exp(-\delta\rho)$. A link can satisfy the requirement if

$$\delta \leqslant \frac{1}{2}\frac{\rho}{\sigma_\Sigma^2}$$

Thus, if we plug in $\rho = \log(M/\mu)$, we ge

$$\log\mu + 2\delta\sigma_\Sigma^2 \leqslant \log M \tag{5}$$

What we achieved by deriving the condition given in (5) is that we took complicated probability relations for a channel and translated them into a linear relation involving the first two order statistics of the loss characteristics of the channel and the application parameter δ. If we pick a link satisfying (5), then we can meet the application's requirement, P_{app}.

One way to interpret the condition in (5) is as follows. Suppose the application does not specify any loss probability

[3] The ETX estimate in [3] is $1/\text{E}\left[p_c(T)\right]$. We use Eq. (2) as the expected number of transmissions.

constraint, i.e., $\delta = 0$. The condition turns into a a comparison of μ, the average channel behavior, and M. Thus, the application requirement turns into a condition involving averages only. Now suppose the application has a loss rate requirement, i.e., $\delta > 0$. In that case we need to *overbook* our resources to meet the loss probability target. The amount of spare μ that has to be put aside in order to accommodate fast time scale fluctuations through overbooking is $2\delta\sigma_\Sigma^2$. This way the packet loss probability target is met. As expected, this amount is directly related to the variability, σ_Σ^2, of the channel and the strictness, δ, of the application loss rate requirement.

The first and second terms on the left side of (5) are the expected value and the scaled version of the variance[4] of the log instantaneous number of transmissions, respectively. We compare the sum of these two with log maximum number of transmissions before a packet loss manifests at the higher layer (because the maximum number of link-layer transmissions, M has been exhausted). The unit of the right side is log number of transmissions and so is the unit of the left side. This sum can hence be thought of as the logarithm of the **effective number of transmissions** (*i.e.,* $\log \mathrm{ENT}$) of the link.

For a wireless link, l, we define

$$\begin{aligned}\alpha^{(l)}(\delta) &= \mathrm{E}\left[\log \frac{1}{p_c(T)}\right] + 2\delta \mathrm{var}\left(\log \frac{1}{p_c(T)}\right) \\ &= \mu + 2\delta\sigma_\Sigma^2\end{aligned}$$

as the $\log \mathrm{ENT}$ (and hence $\exp(\alpha(\delta))$ is the ENT). If a link satisfies $P_{\mathrm{app}} \leqslant \exp(-\delta\rho)$, then the $\log \mathrm{ENT}$ for that link is between the expected log instantaneous number of transmissions and the maximum number of link-layer transmissions, M, before a packet loss is observed at the sender's transport layer. As a result, for link l, one can write the following condition:

$$P_{\mathrm{app}} \leqslant \exp(-\delta\rho) \quad \Rightarrow \quad \alpha^{(l)}(\delta) \leqslant \log M \qquad (6)$$

3.3 Geometric Interpretation of QAR Metrics

To shed some light on the ENT metric and its relation to other quality-aware metrics, we now present a geometric interpretation of ENT. Insights from this interpretation may lead to a better understanding of what ENT signifies in practice and will help us view different QAR algorithms in a unified manner.

Let us represent a wireless link by two parameters, $\log(\mu/M)$ ($= -\rho$) and σ_Σ. Each link corresponds to a point in the coordinate space $(\sigma_\Sigma, \log(\mu/M))$ as illustrated in Fig. 5. Thus, the point with the lowest ordinate value is the one that minimizes the expected number of transmissions. Such links will be preferred by routing algorithms that employ ρ as the link cost metric (*e.g.,* ETX).

The set of points that satisfy $P_{\mathrm{app}} = \exp(-\delta\rho)$ are on the parabola $\alpha(\delta) = \log M$ as shown in Fig. 6. Thus, the points outside of the shaded region fail to satisfy $P_{\mathrm{app}} \leqslant \exp(-\delta\rho)$. The shaded region can therefore be regarded as a *feasible region*. Suppose we want our routing algorithm to select the links that not only minimize the expected number of transmissions, but also keep the loss probability smaller than $\exp(-\delta\rho)$. Then the algorithm should pick the link with the smallest ordinate value among the points in the feasible region. We describe this clearly in the first algorithm we give at the end of the section.

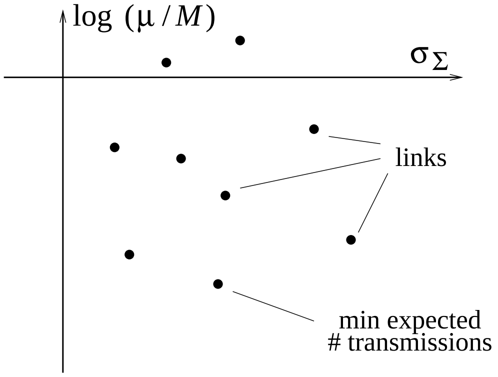

Figure 5: Each point in $(\sigma_\Sigma, \log(\mu/M)$ coordinate space represents a link.

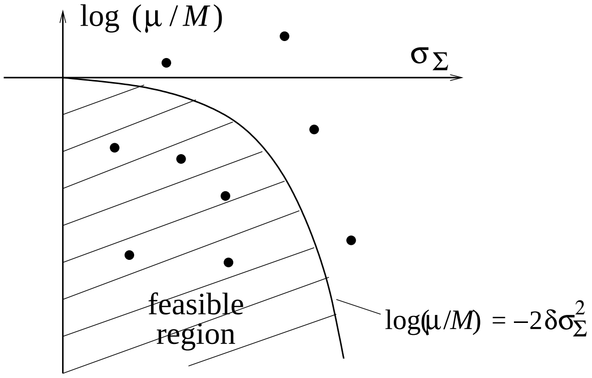

Figure 6: The points in the feasible region satisfy $P_{\mathrm{app}} \leqslant \exp(-\delta\rho)$.

From Eq. (4),

$$\log \mathrm{P}\left(\frac{1}{p_c(T)} \geqslant M\right) \approx -\frac{1}{2}\left(\frac{\log(\mu/M)}{\sigma_\Sigma}\right)^2$$

Since for any given point, the slope of the line connecting the origin to that point is $\log(\mu/M)/\sigma_\Sigma$, points with larger $|\log(\mu/M)/\sigma_\Sigma|$ has lower loss probability. For instance, in Fig. 7, channel l has a lower loss probability than channel l'. If the objective is to minimize the probability of loss, then the routing algorithm will choose points with large $|\log(\mu/M)/\sigma_\Sigma|$ ratios.

Finally, let us evaluate the vertical distance, $D^{(l)}$, between any admissible point, $l : (\sigma_\Sigma, \log(\mu/M))$ and the boundary of the feasible region. As illustrated in Fig. 8,

$$\begin{aligned}D^{(l)} &= -\log\left(\frac{\mu}{M}\right) - 2\delta\sigma_\Sigma^2 \\ &= \log M - \alpha^{(l)}(\delta) \qquad (7)\end{aligned}$$

[4] As S grows, σ_Σ^2 approaches index of dispersion of the process $p_{B,t}$

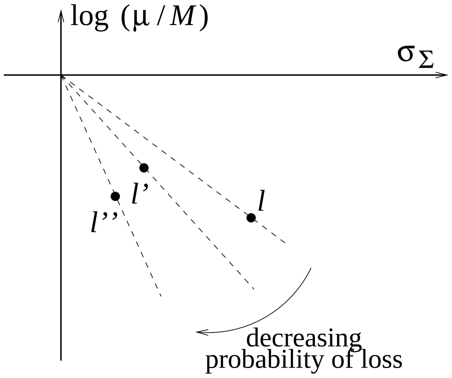

Figure 7: The slopes of the dashed lines are representatives of the loss probability.

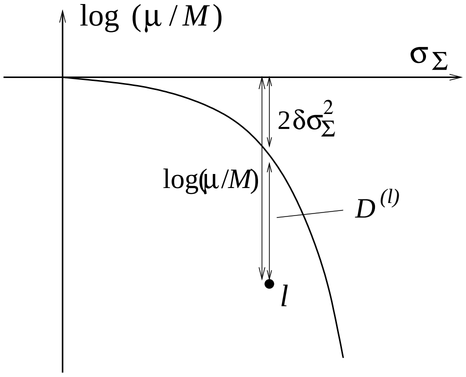

Figure 8: The vertical distance between a point and the boundary of the feasible region is $\log M$ − log **ENT.**

Hence, the link that maximizes the vertical distance to the boundary of the feasible region is the one that minimizes the ENT. This means, given an increase in the expected number of transmissions, the link with a small ENT is more likely to remain in the admissible region. Thus, if the objective is robustness with respect to the changes in the expected number of transmissions, the routing algorithm will choose points with smaller ENT.

3.4 ENT-based Routing Protocols

Based on the insights we built in the previous section, we will construct three routing algorithms based on ENT and loss rates. In terms of dissemination of routing information, these algorithms are very similar to the ETX algorithm. Indeed the only difference is that the parameter σ_Σ is propagated as well as the expected number of transmissions.

There are two points we need to emphasize before we give these algorithms. The first one is that since the ENT is a function of the application parameter, δ, an optimal path between two nodes may differ from one application to another. The second is that we assume the process $p_{B,t}$ to be uncorrelated for different links in a network (this is observed in [3]). Thus, σ_Σ^2 is additive over the different links on a path between two nodes.

Algorithm 1: For each link, compute the log ENT. Compare against $\log M$. Assign a cost of ∞ to the links that have log ENT $> \log M$ and assign a cost of ETX to the others. Between any given pair of nodes use the path that minimizes the total cost.

This algorithm focuses only on the feasible links, i.e., the ones that satisfy the application loss reqirement, P_{app}. It picks those with the minimum ETX among those. This involves only a minimal modification to ETX (given the link parameters).

Algorithm 2: For each link, compute the log ENT. Compare against $\log M$. Assign a cost of ∞ to the links that have log ENT $> \log M$ and assign a cost of log ENT to the others. Between any given pair of nodes use the path that minimizes the maximum ENT over all the links that belong to the path.

This algorithm also focuses on the feasible links. It picks the path which is most robust with respect to the changes in the channel. I.e., it picks the paths for which the links tend to remain in the feasible region longer than others.

Algorithm 3: For each link, compute the log ENT. Compare against $\log M$. Assign a cost of ∞ to the links that have log ENT $> \log M$ and assign a cost of $-\log(1 - P_{\text{loss}})$ to the others. Between any given pair of nodes use the path that minimizes the total cost.

This too focuses on the feasible links. Between any given pair of nodes, it picks the path with the minimum overall probability of packet loss over all the links.

One can notice that we can arbitrarily add extra constraints to the problem (such as a constraint on ETX for example) depending on the requirements of the application and solve the problems without much extra complexity. This is due to the simplicity of the way (i.e., the ENT) we handle the high dimensionality of the problem.

4. Channel Estimation

So far, we have assumed that the channel parameters μ, σ_Σ are available to us at any point in time. In reality, QAR metrics depend on the estimation of the parameters that represent the channel characteristics. We have to estimate these parameters based on the measurements taken from the channel.

There are two types of measurements depending on how much physical layer information we have. In the first case, $p_{B,t}$ may be available to the estimator. This may be partially reasonable if some detector reliability information available from the physical layer can be accessed by the estimator. The signal to noise ratio (SNR) is an example of such information since there is a direct relation between bit (or symbol) error probability and SNR. However, in most systems, values of $p_{B,t}$ are not available for the use of the estimator. In such cases, active or passive probing may be used. Probe packets are then examined for bit errors. Thus, instead of individual samples of $p_{B,t}$, the estimator has the knowledge of the outcome of sequence of indicator random variables, $1_{B,t}$.

$$1_{B,t} = \begin{cases} 1 & \text{w.p. } p_{B,t} \\ 0 & \text{w.p. } 1 - p_{B,t} \end{cases}$$

If we extract $1_{B,t}$ from the probe packets, we can examine the positions of bit errors and estimate how they are correlated.

We study estimators based on different types of information available. Before we present our estimators, let us give some fundamental assumptions we make about the dynamics and the time scales of the network.

4.1 Time Scales

Let T_r be the time scale that the changes in routes occur. Namely, once a path is assigned for a pair of nodes it is kept unchanged for a duration of T_r time units. Routing time scale is typically in the order tens of seconds (i.e., hundreds of packets), so the network updates optimal routes once every T_r seconds. This is an exogenous parameter of the system and picked independently of the network parameters such as μ, σ_Σ.

As we mentioned before, the process $p_{B,t}$ fluctuates over multiple time scales. A significant portion of the power of this process is in frequencies $f > T_r^{-1}$. Thus, the channels vary[5] significantly within a routing time scale. A metric that is based on the average link behavior can only capture the effects of the components that do not exhibit a significant variation between routing updates. Thus, the channel estimation filters for such metrics as ETX need to have cutoff frequencies at $f \approx T_r^{-1}$.

Designing such low pass filters has received much attention in past work on channel estimation for routing metrics. Various filters from the simple moving average ([3]) and exponentially weighted moving averaging (EWMA) to more sophisticated non-linear filtering ([9, 15, 16]) have been proposed in the literature. In [9] and [15], the EWMA parameters are adjusted according to the channel conditions to find a balance between *agility* and *stability*.

By incorporating the variance, we take the faster time scale components, $f > T_r^{-1}$, in addition to the slowly varying ones. The high frequency components are used to estimate the amount of "spare" ETX that has to be put aside (the variance term in the ENT) in order to accommodate fast time scale fluctuations through "overbooking." This way the packet loss probability target is met. Indeed, this constitutes the main idea behind ENT routing metric. Without taking into account the high frequency terms as well as low frequencies, it is not possible to meet the loss rate requirement, P_{app}.

Recall that we assumed that $p_{B,t}$ is stationary for each link. Now we relax this assumption. Let T_c be the time scale over which the statistics of a link remain unchanged. I.e., μ, σ_Σ^2 exhibit either little or no change in any given consecutive T_c time slots. We assume that $T_c \gg T_r$. We also assume that the estimators have an idea of the value of T_c. The reason why we make this assumption is that we need accurate estimates of the sample statistics (sample mean, autocovariance) in our estimators. Note that, this assumption does not imply that channels do not vary in time scales less than T_c. Indeed, a significant portion of their variations occur in time scales $< T_c$. It just means that the statistics are unchanged in T_c.

4.2 Estimating the ENT

In this section we discuss issues associated with estimation of ENT and describe our ENT estimator. The process, $p_{B,t}$ or $1_{B,t}$ are measured using probe packets. Each node broadcasts 8-10 probe packets approximately once every second and extracts the sequence of $p_{B,t}$ or $1_{B,t}$ depending on whether there is access to the physical layer or not.

To eventually estimate the ENT using these samples, we first need to estimate the sample mean and the autocovariance function. We use a moving average of duration T_c seconds which includes T_c sets of probe packets since a probe packet is send every second. Note that we will have an accurate estimate of the mean and the samples of the autocovariance function and these samples are separated by a second (sets of probe packets are separated by that a second). If we use $1_{B,t}$, the sample mean, $\hat{\mu}_{1_B}$ and the autocovariance estimate, $\hat{K}_{1_B}(t)$ converge to the sample mean, $\bar{p}_B$ and autocovariance, $K_{p_B}(t)$ of $p_{B,t}$ respectively. Note however that, σ_Σ^2 cannot be estimated using the samples of $1_{B,t}$.

At the beginning of each routing period (once every T_r), we use the most recent set of probe packets to estimate ENT to be used in each routing period. Thus, we have the following prediction problem at the beginning of each routing period. We take the beginning of the routing period as $t = 0$. Given the values of $p_{B,s}$ or $1_{B,s}$, $0 < s \leqslant \tau$ where τ is the duration of a probe period (8-10 packets) we would like to estimate$\Sigma_k = \sum_{s=\tau+1+kS}^{\tau+(k+1)S} p_{B,s}$ for all k, $0 \leqslant k \leqslant T_r/S$ where T_r/S is the routing time scale given in terms of number of packets. Then, using these individual Σ estimates, we will assign an ENT for the link to be held in the entire routing period. We emphasize that even though we have T_r sets of probe packets within a routing period, each path will remain fixed over the entire routing period.

Before giving our estimator, to give an understanding on the agility-stability trade-off[6], we present a simple estimator. Suppose we have access to the physical layer information and thus have samples of $p_{B,t}$ available to us. From the long observation windows, we have accurate estimates of the sample mean, $\bar{\Sigma} = S\bar{p}_B$ and the sample variance, σ_Σ^2 of Σ_k. For link l, we can use the estimate

$$\hat{\alpha}(\delta) = \bar{\Sigma} + 2\delta\sigma_\Sigma^2 \tag{8}$$

for all the routing periods within the entire T_c. This ENT estimate is very accurate, but it changes once every T_c time slots. In many networks this simple estimator may be preferable since it captures the statistical changes in the channel and it does not even need the long observation window.

[5] This is the main motivation for using the channel variations in our metrics

[6] In the estimation literature this is also known as the error-resolution trade-off

However, there is significant power in components of $p_{B,t}$ that vary in time scales T, $T_r < T < T_c$. If we would like to capture such variations, we need to use more *agile* estimators that use the local observations as well as the long observation window. The question is can we do this without sacrificing the accuracy. The answer is yes. Next, we will build an "optimal" estimator that takes the estimation errors into consideration in constructing the ENT estimate.

4.2.1 Optimal Linear Estimator

We will construct the linear least squares (LLS) estimator to predict Σ_k for all k, $0 \leqslant k \leqslant T_r/S$. Then, we will combine these individual estimates to get a single ENT estimate for the routing period.

The autocovariance function determines the coefficients of the LLS estimator which is the filter that minimizes the mean squared error. We build the estimator for the case where $p_{B,t}$ is available which can be replicated identically for the case where $1_{B,t}$ is available.

As described, we have the accurate estimates of μ and $K_{\vec{p}_B}$ of $\vec{p}_B = [p_{B,1}\ p_{B,2}\ \cdots\ p_{B,\tau}]'$. Hence, we shall use $\bar{p}_B$ instead of $\hat{\mu}$ and $K_{\vec{p}_B}$ instead of $\hat{K}_{1_B}(t)$. Let $\vec{K}_{\Sigma_k \vec{p}_B} = \mathrm{cov}\,(\Sigma_k, \vec{p}_B)$ which is a τ dimensional vector its s entry is

$$\vec{K}_{\Sigma_k \vec{p}_B}(s) = \sum_{l=\tau+kS+1}^{\tau+(k+1)S} K_{p_B}(l-s)$$

The coefficients of the LLS estimator for Σ_k can then be found as (see [17]):

$$\hat{\Sigma}_k^{\text{lls}} = S\bar{p}_B + \vec{K}_{\Sigma_k \vec{p}_B} K_{\vec{p}_B}^{-1} (\vec{p}_B - \bar{p}_B) \qquad (9)$$

and the mean squared error, ϵ_k, for the LLS estimate is:

$$\begin{aligned} \epsilon_k^{\text{lls}} &= \mathrm{E}\left[(\Sigma_k - \hat{\Sigma}_k^{\text{lls}})^2\right] && (10) \\ &= \sigma_\Sigma^2 - \vec{K}_{\Sigma_k \vec{p}_B} K_{\vec{p}_B}^{-1} \vec{K}'_{\Sigma_k \vec{p}_B} && (11) \end{aligned}$$

Note that the second term in Eq. (11) is the decrease in the error variance due to using LLS estimator instead of the plain mean estimator, (8). Let us focus in the definition of the error variance given in (10). It gives us the mean square deviation of Σ_k from its estimate, $\hat{\Sigma}_k^{\text{lls}}$. Thus, the estimated log ENT, $\hat{\alpha}(\delta, \tau + kS)$ at time τ for the actual ENT of the link at time $\tau + kS$ is

$$\hat{\alpha}(\delta, \tau + kS) = \hat{\Sigma}_k^{\text{lls}} + 2\delta\epsilon_k^{\text{lls}} \qquad (12)$$

An interesting point to emphasize here is that the variance term of the log ENT estimate is due to the variance of the estimation error. This may look counter intuitive since σ_Σ^2 and ϵ_k are different quantities. The reason why we use the latter in place of the former becomes clearer when we consider the two extreme cases. First suppose we can estimate Σ_k perfectly, i.e., $\epsilon_k = 0$. Then, $\hat{\Sigma}_k = \Sigma_k$ with probability 1, which means Σ_k is no longer random. Therefore,

$$\hat{\alpha}(\delta, \tau + kS) = \hat{\Sigma}_k$$

in which there is no δ term. Second, suppose $\vec{p}_B$ and Σ_k are uncorrelated. Then,

$$\hat{\alpha}(\delta, \tau + kS) = S\bar{p}_B + 2\delta\sigma_\Sigma^2$$

since $\epsilon_k = \sigma_\Sigma^2$. Thus, the ENT decreases with decreasing estimation error. This suggests that if ENT is used as the routing metric, the best way home is the way you know!

4.2.2 Combining Individual ENT Estimates

Now we know how to estimate the ENT for individual time periods, $0 \leqslant k \leqslant T_r/S$. We have to come up with a single metric to find the best route for the entire routing period. So, let us discuss how to combine these ENT estimates, $\hat{\alpha}(\delta, \tau + kS)$ to find that single metric. It is proved in Appendix B that the "best" log ENT estimate for the period $(\tau, \tau + T_r)$ is

$$\hat{\alpha}(\delta) = \frac{1}{T_r/S} \sum_{k=1}^{T_r/S} \left[\hat{\alpha}(\delta, \tau + kS) + 2\delta(\hat{\Sigma} - \hat{\Sigma}_k)^2\right] \qquad (13)$$

where

$$\hat{\Sigma} = \frac{1}{T_r/S} \sum_{k=1}^{T_r/S} \hat{\Sigma}_k$$

The overall ENT estimate given in (13) is the average of the individual ENT estimates for all $k \leqslant T_r/S$ increased by the extra error for using $\hat{\Sigma}$ instead of $\hat{\Sigma}_k$ for each individual packet.

We can build the LLS estimator for $\Sigma(k)$ based on the observations, $1_{B,s}$, $0 < s \leqslant \tau$, using the matrix $K_{\vec{1}_B}$ instead of $K_{\vec{p}_B}$ in Eq. (9). One can observe that

$$K_{\vec{1}_B} = K_{\vec{p}_B} + \left(\bar{p}_B - \mathrm{E}\left[p_{B,t}^2\right]\right) I \qquad (14)$$

where I is the identity matrix. The eigenvectors of $K_{\vec{1}_B}$ and $K_{\vec{p}_B}$ are identical whereas each eigenvalue of $K_{\vec{1}_B}$ is $\bar{p}_B - \mathrm{E}\left[p_{B,t}^2\right]$ higher than the corresponding eigenvalue of $K_{\vec{p}_B}$. The increase in eigenvalues causes an increase in the error variance, ϵ_k. Indeed, the negative term in Eq. (11) decreases inversely proportional to the increase in the eigenvalues.

4.3 Examples

In this section, illustrate the channel estimation tools we developed so far on one real world and two toy examples. We shall also give plots for the traditional EWMA filter and the *modified* EWMA filter which was developed by us and compare the results of these two filters. A detailed treatment of these filters can be found in Appendix C. In the first example, we assume that the autocovariance of the loss rate consists of a single exponentially decaying term, i.e., the $p_{B,t}$ process varies in single time scale. The second example illustrates the scenario with two terms that are exponentially decaying and one term which decays as $t^{-\beta}$, $0 < \beta < 1$ so that the resulting process varies in two time scales and it is long range dependent. The third example considers the trace of $1_{B,t}$ taken from a packet exchange between two nodes separated by 0.55 miles. In these three examples we evaluate ENT estimates and study how different filters perform depending on their system parameters such as the ratio of probing time scale to routing time scale, rate of decay of the autocovariance function. Depending on the results we have, we reach certain conclusions on how to do channel estimation.

4.3.1 Single Time Scale

In this example, the autocovariance function of $p_{B,t}$ decays exponentially and it varies over a single time scale. Indeed, the autocovariance,

$$K_{p_B}(t) = \sigma_{p_B}^2 e^{-\beta t/S}$$

In the case where $p_{B,t},\ t \leqslant \tau$ is known, the LLS estimator coefficients are all zero except for the coefficient of $p_{B,\tau}$ term.This means, only the most recently observed sample is relevant for all the predictions in the future. Also, this term only has limited impact for estimating farther points since the coefficient of this most recently observed term in the estimate decays exponentially with the decay rate β. I.e., the LLS estimate converges to the sample mean estimate exponentially fast!

Another thing to point out is that just one probe packet is sufficient in this case since we use just the most recently observed sample, $p_{B,\tau}$. Not however that the long probe sequence once every T_c is necessary to estimate the mean and the autocovariance.

The case where we observe $1_{B,t}$ is very different. This time the LLS estimator uses a larger set of observations. In what follows we assume $S\bar{p}_B = 1$, $\sigma_{p_B}^2 = 0.01$ and $\beta = 0.01$. In Fig. 9 we illustrate the LLS estimator coefficients for $\hat{\Sigma}_1^{\text{lls}}$ for $\tau/S = 2,\ 5$ and 10 packets. The bit position is inverted since it goes from the recent past to further in the past. Thus, position 0 in the graph represents the coefficient for $p_{B,\tau}$. Packet sizes are normalized to $S = 10$ bits. Error variance[7] for these three cases are 0.407, 0.361 and 0.359 respectively. It does not decrease further significantly with increased probe size (the decrease in the error variance is approximately 10^{-5} when τ is increased to 100 packets from 10 packets.). Note that, if $p_{B,t}$ is known, $\epsilon_1^{\text{lls}} = 0.055$ which is much less than $\epsilon_1^{\text{lls}} =$ of the case where $1_{B,t}$ is observed. Also, as the decay rate of the autocovariance function increases, $\hat{\Sigma}_k^{\text{lls}}$ looks more and more like the simple moving average estimator.

In Fig. 10, we plot the estimator coefficients and associated error variances for $\hat{\Sigma}_{T_r/S}^{\text{lls}}$ for different values of τ/T_r. We took $\tau/S = 10$ and normalized the packet size to $S = 10$. We found that $\epsilon_{T_r/S} = 0.913,\ 0.988$ and 0.999 for $T_r/\tau = 1,\ 2$ and 5 respectively. $\vec{p}_B$ and $\Sigma_{T_r/S}$ become almost uncorrelated for $T_r/\tau > 3$. This example illustrates the scenario where the observation window is small, i.e., the probe packet sequence is short compared to the routing time scale. As a side note, these values are also ϵ_{10}, ϵ_{20} and ϵ_{50} for the case where T_r/τ is fixed at 5.

The error variances for the case where $\vec{p}_B$ is observed are 0.872, 0.983 and 0.999 for $T_r/\tau = 1,\ 2$ and 5 respectively. The interesting point is that the difference between the error variances of the LLS estimates when $p_{B,t}$ is observed and when $1_{B,t}$ is observed becomes insignificant for increasing T_r/τ.

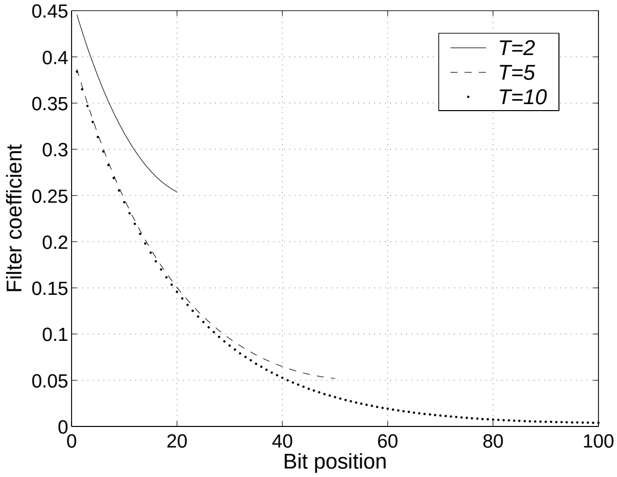

Figure 9: Estimator coefficients for $\hat{\Sigma}_1^{\text{lls}}$ as a function of the bit position of the probe bits for varying τ. We normalized the packet size to $S = 10$.

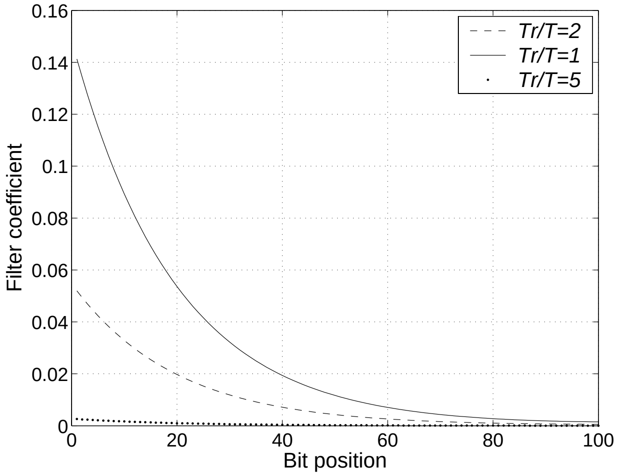

Figure 10: Estimator coefficients for $\hat{\Sigma}_{T_r/S}^{\text{lls}}$ as a function of the bit position of the probe bits for varying T_r/τ. We normalized the packet size to $S = 10$.

[7] We give the error variance as a fraction of σ_Σ^2. If we say the error variance is 0.5, we mean that the variance of error is $\sigma_\Sigma^2/2$. From now on, when we say error variance, we mean *relative* error variance.

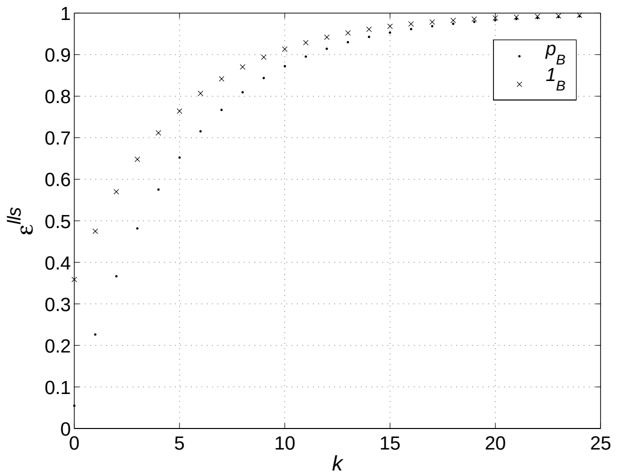

Figure 11: Estimation error variance, $\epsilon_k^{\mathrm{lls}}$, for $\tau/S = 10$ and $T_r/\tau = 2$.

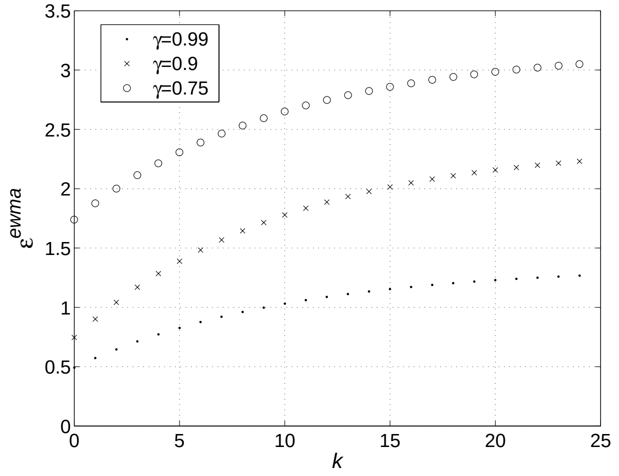

Figure 12: Estimation error variance, $\epsilon_k^{\mathrm{ewma}}$, of the generic EWMA estimator for $\tau/S = 10$ and $T_r/\tau = 2$.

In Fig. 11, $\epsilon_k^{\mathrm{lls}}$, $k \leqslant 25$ is illustrated for $\tau/S = 10$ packets for both $\vec{p}_B$ and $\vec{1}_B$.

Fig. 12 illustrates the error variance, $\epsilon_k^{\mathrm{ewma}}$, of $\hat{\Sigma}_k^{\mathrm{ewma}}$ with the generic EWMA estimator for $\gamma = 0.99$, 0.9 and 0.75. The observation is $\vec{1}_B$ and we keep T_r/τ fixed at 2.5. It can be seen that the EWMA filter performs better than MA (since $\epsilon_k^{\mathrm{ma}} = 1$ for all k) only for a couple of k values. Beyond that, it has a higher variance than 1 which makes it inappropriate for loss rate estimation in this example.

To illustrate how much gain we get by estimating the second order statistics of the loss rate process we also considered the modified EWMA estimator. The modified EWMA estimator is illustrated for the same parameters in Fig. 13. Also, the error variance of the LLS estimator is plotted on the same graph. For a good range of γ values, the modified EWMA estimator performs very similar to the LLS estimator. As mentioned, it is much simpler to implement since it does not require the matrix inversion as the LLS estimator.

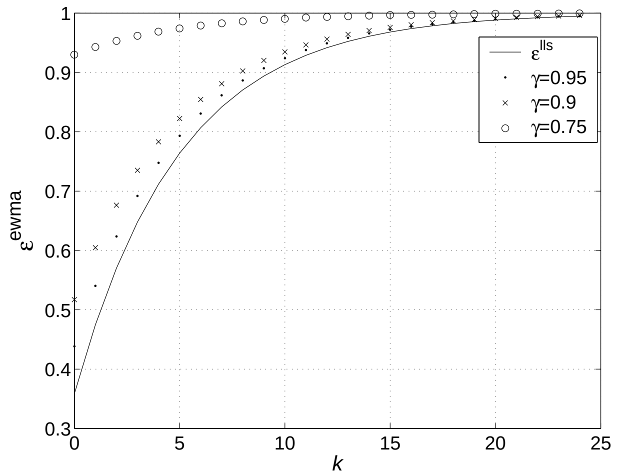

Figure 13: Estimation error variance, $\epsilon_k^{\mathrm{ewma}}$, of the modified EWMA estimator for $\tau/S = 10$ and $T_r/\tau = 2$.

4.3.2 Multiple Time Scales

In this example, the autocovariance function of the $p_{B,t}$ process has two exponentially decaying components and a slowly decaying term which makes the $p_{B,t}$ process long range dependent. In [8, 10], the authors imply that the packet loss probability exhibits similar behavior by saying that the packet loss bursts are small with high probability but occasionally very long bursts are observed.

$$\begin{aligned} K_{p_B}(t) &= \sigma_{p_B}^2 \Big[C_1 e^{-\beta_1 t/S} + C_2 e^{-\beta_2 t/S} \\ &+ (1 - C_1 - C_2)(t/S)^{-\beta_3} \Big] \end{aligned}$$

$S\bar{p}_B = 1$ and $\sigma_{p_B}^2 = 0.01$. We take $\beta_1 = 0.01 \ll \beta_2 = 0.1$ and $C_1 = C_2 = 0.45$. Thus, these components have equal power and the first one is more dominant in low frequencies and the second term is more dominant in higher frequencies. We take $\beta_3 = 0.5$, and thus the third term has infinite power in DC! This means that the $p_{B,t}$ is long range dependent. In [8], the authors imply that packet loss probability exhibits similar behavior by saying that the packet loss bursts are small with high probability but occasionally very long bursts are observed.

LLS estimator coefficients are illustrated in Fig. 14 for $\hat{\Sigma}_3^{\mathrm{lls}}$ (the third packet after the probe sequence) for $\tau/S = 2$, 5 and 10 packets. The bit position is inverted since it goes from the recent past to further in the past. Thus, position 0 in the graph represents $t = \tau$. Error variances are 0.821, 0.811, 0.81 respectively. It does not decrease further significantly with increased probe size.

In the case where $p_{B,t}$, $t \leqslant \tau$ is known, the LLS estimator has a very high coefficient for the $p_{B,\tau}$ term as was the case in the single time scale scenario. However, this time other coefficients are non-zero. More interestingly, there is a spike at $p_{B,0}$, i.e., the coefficient of $p_{B,0}$ is much higher than its

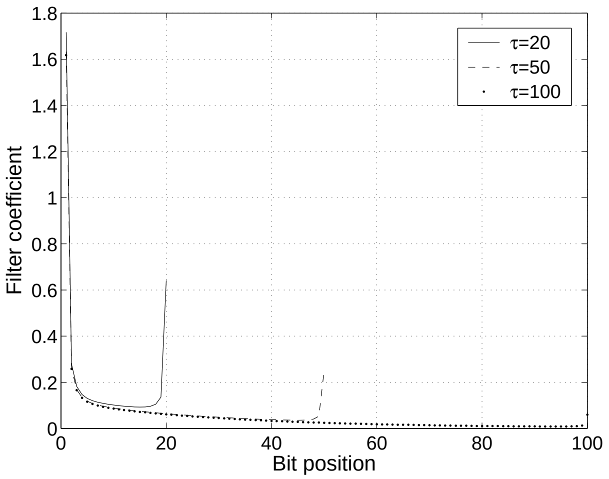

Figure 14: Estimator coefficients for $\hat{\Sigma}_3^{\mathrm{lls}}$ for varying τ. We normalized the packet size to $S = 10$.

close neighbors. At the beginning this looks surprising since $p_{B,0}$ has the lowest covariance with Σ_k among all $p_{B,t}$. The explanation for this "anomaly" is that $p_{B,0}$ is the representative for all the points that are not observed. For instance, suppose we would like to estimate $p_{B,s}$ for some $s < 0$ using $p_{B,t},\ 0 \leqslant t \leqslant \tau$. The LLS estimator will put a very high weight on $p_{B,0}$ compared to all other observations. This means, among $p_{B,t},\ 0 \leqslant t \leqslant \tau$, $p_{B,0}$ carries significantly higher information about the entire past than all the other points. Therefore, when estimating Σ_k, the LLS filter puts the weight of all unobserved points (i.e., $p_{B,t},\ t < 0$) onto $p_{B,0}$ as their representative.

In Fig. 15 we illustrate the LLS estimator coefficients for $\hat{\Sigma}_1^{\mathrm{lls}}$ for $\tau/S = 2,\ 5$ and 10 packets. Packet sizes are normalized to 10 bits. Error variance for these three cases are 0.893, 0.869 and 0.863 respectively. A more smoothed version of the spiky behavior that we had in the $p_{B,t}$ case for the earlier probe bits is observed in this case. The reason is the same.

For this example, instead of plain error variances, we study how $\epsilon_k^{\mathrm{lls}}$ approaches to 1 as k increases. This gives us an understanding on the effect of the long range dependent term of the autocovariance on the estimation error. The *complementary* error variance, $1 - \epsilon_k^{\mathrm{lls}}$ is given in Fig. 16 for $T_r/\tau = 25$ and $\tau/S = 10$ packets. Beyond $k = 100$, the impact of the exponentially decaying terms are negligible. After this point, the decay in the complementary error variance is as $t^{-2\beta_3} = t^{-1}$. Thus, unlike the first example, in systems with long range dependent bit error probabilities, the estimates may carry non-negligible information even far away from the observation points.

Fig. 12 illustrates the error variance, $\epsilon_k^{\mathrm{ewma}}$, of $\hat{\Sigma}_k^{\mathrm{ewma}}$ with the generic EWMA estimator for $\gamma = 0.99,\ 0.95$ and 0.9. The observation is $\vec{1}_B$ and we keep T_r/τ fixed at 2.5. The generic EWMA estimator performed similar to the previous example which makes the MA filter more appealing.

The error variance for the modified EWMA estimator is il-

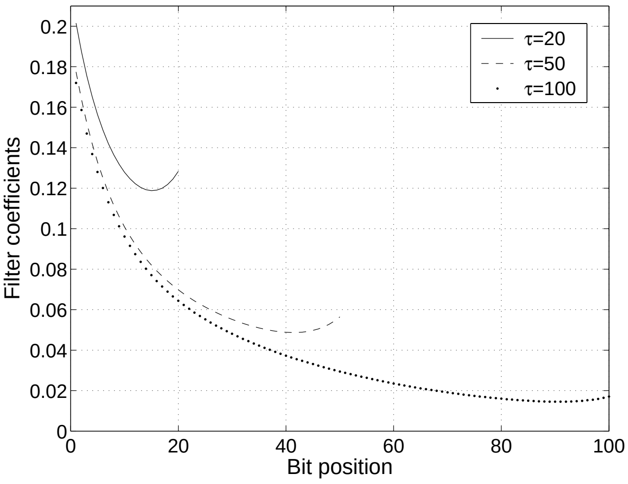

Figure 15: Estimator coefficients for $\hat{\Sigma}_1^{\mathrm{lls}}$ as a function of the bit position of the probe bits for varying τ. We normalized the packet size to $S = 10$.

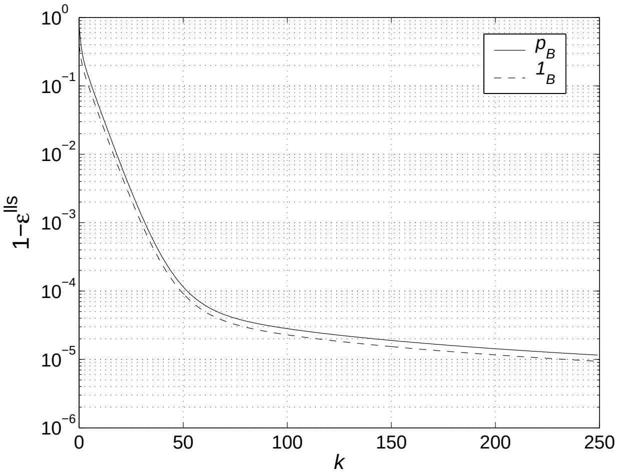

Figure 16: The complementary error variance decays as $t^{2\beta_3} = t^{-1}$ after the exponentially decaying components become insignificant.

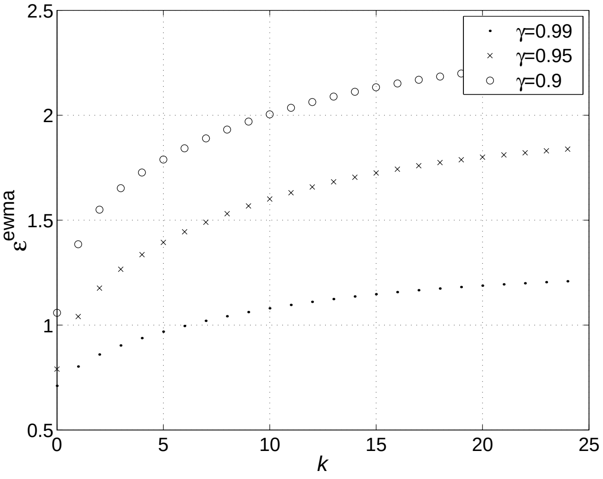

Figure 17: Estimation error variance, ϵ_k^{ewma}, of the generic EWMA estimator for $\tau/S = 10$ and $T_r/\tau = 2$.

lustrated for a range of parameters in Fig. 18. We take $T_r/S = 1000$ packets. Also, the error variance of the LLS estimator is plotted on the same graph. For a range of γ values ($\gamma \geqslant 0.81$), the complementary error variance decays as t^{-1}. There is critical point $\gamma \approx 0.81$ in this example) beyond which there exists a $k < \infty$ such that $\epsilon_k^{(ewma)} > 1$. We believe such a point exists for all possible calibrations of the modified EWMA estimator. We must reduce the estimator coefficient for $\hat{\Sigma}_k^{\text{lls}}$ further while keeping it proportional to $K_{\vec{1}_B \Sigma_k}$. The optimal calibration is beyond the scope of this paper.

4.3.3 Real Trace

In this example we use the trace given in Fig. 3. This trace consists of 194 packets of size $S = 1392$ bytes. Since the position of individual byte losses are not available to us, we use a different approach here. We use the number of bytes lost,

$$N_k = \sum_{t=S(k-1)+1}^{Sk} 1_{B,t}$$

In this example, we use $N_k,\ k \leqslant \tau$ to estimate $\Sigma_k,\ k > \tau$.

First we estimate the statistics of N_k. Mean error probability, $\bar{N}_k = 225.922$. Note that this is also $\mathrm{E}\left[\Sigma_k\right]$. The estimated autocovariance function, $\hat{K}_N(t)$ of N_k is given in Fig. 19. The interesting thing about the autocovariance function is that it has an envelope with an amplitude which increases, i.e., the absolute covariance tends to increase with increasing t. The sample variance of N_k is $\sigma_\Sigma^2 \approx 2.42 \times 10^5$. This shows that the loss probability may depend highly on the variance rather than the mean of the probability of loss. The power spectrum, $S_N(\omega)$ of N_k was given in Fig. 4.

After estimating the statistics, we use the first $\tau = 50$ samples of N_k as probe packets and find the LLS estimate, $\hat{N}_k^{\text{lls}}$ for $51 \leqslant k \leqslant 150$. The theoretical error variance, ϵ_k^{lls} of the LLS estimator is given in Fig. 20. The estimation error for the packets close to the probe packets are around 0.8

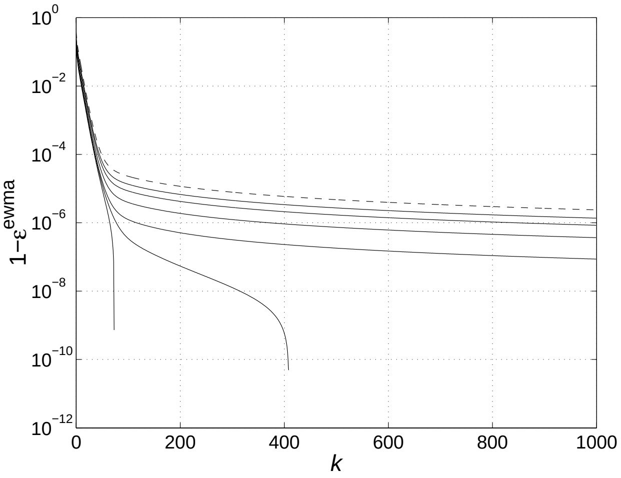

Figure 18: The dashed curve is the complementary error for the LLS estimator and the solid curves are the complementary error variance for the modified EWMA estimator for $\gamma = 0.95,\ 0.9,\ 0.85,\ 0.82,\ 0.81$ and 0.8 from top to bottom. The complementary error variance decays with $t^{2\beta_3} = t^{-1}$ for $\gamma > 0.81$. If reduced further, for some $k < \infty$, $\epsilon_k > 1$.

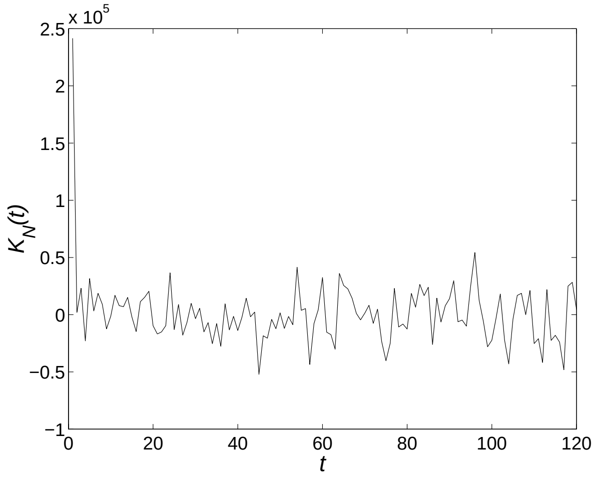

Figure 19: The estimated autocovariance, $\hat{K}_N(t)$ of N_k for $t < 120$.

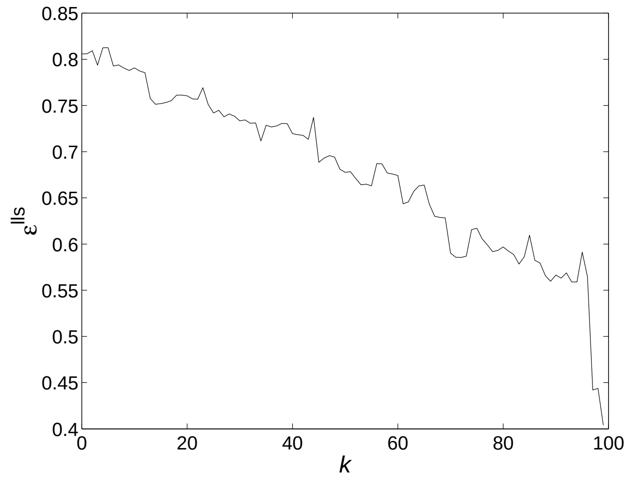

Figure 20: The theoretical error variance, ϵ_k^{lls}, with the LLS estimate.

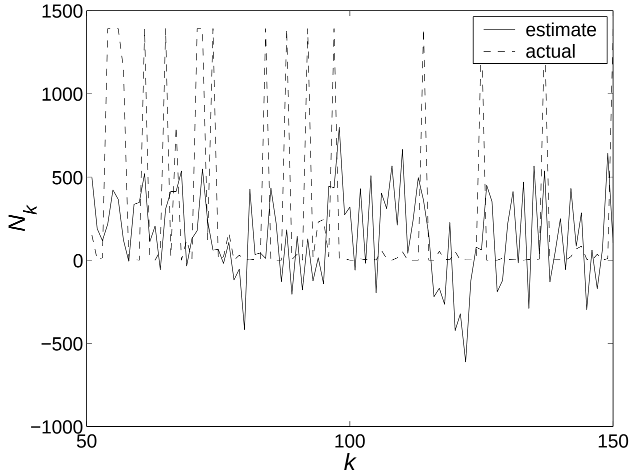

Figure 21: The actual trace and the estimate.

and decreases almost linearly for the subsequent packets in $51 \leqslant k \leqslant 150$ to around 0.4. This is due to the extraordinary nature of the autocovariance function.

The estimates, $\hat{N}_k^{\text{lls}}$ and the actual values of N_k for $51 \leqslant k \leqslant 150$ are illustrated in Fig. 21. The normalized mean squared error for this case is, 1.21 which actually is worse than that of MA. It is related to the high estimation error that we have in estimating the second order statistics and the high power of N_k that is confined in the higher frequencies.

Finally, we use the first $\tau = 50$ samples as probes and use the modified EWMA with $\gamma = 0.99$ (very close to MA) to estimate N_k, $51 \leqslant k \leqslant 150$. The theoretical error variance, ϵ_k^{ewma} of the EWMA estimator is given in Fig. 22.

4.4 Channel Estimation Summary

A number of implications of these results we had can be listed as follows.

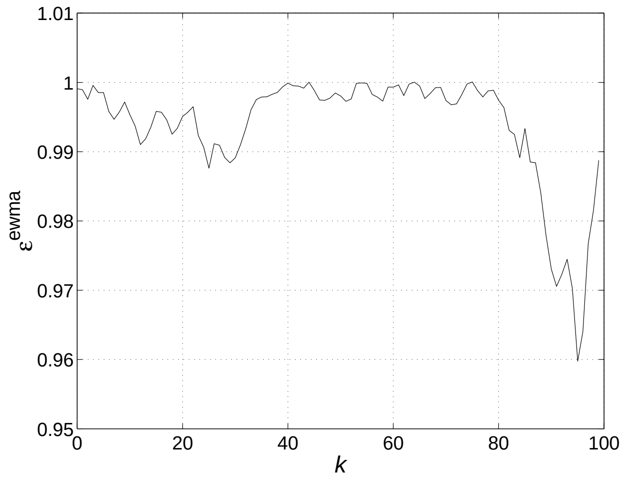

Figure 22: The theoretical error variance, ϵ_k^{ewma}, with the modified EWMA estimate.

- The main estimator we use is the LLS estimator. Almost all the estimators that has a reasonable performance depend on the first two order statistics of the $p_{B,t}$ process. Thus, estimating the first two order statistics is of crucial importance for channel prediction.

- We illustrated that the estimation error variance for the case where we observe $1_{B,t}$ is reasonably close to that where we observe $p_{B,t}$. Note, however that this depends on the assumption that we can measure the first two order statistics in both cases, which is much easier if we have $p_{B,t}$ compared to having $1_{B,t}$.

- The generic EWMA estimation may perform poorly in the time scales of order a few packets. We modified the generic EWMA to improve its performance highly, but the modified version depends on the knowledge of the first two order statistics of the channel. In the time time scales of interest, simple MA estimate is much more appealing due to its simplicity and it does not require the knowledge of the mean and the autocovariance of $p_{B,t}$.

- The variance of $p_{B,t}$ plays an significant role in system performance and it should be taken into consideration in quality aware routing protocols.

- The motivation of ENT is not just measuring the higher layer losses. It is also the fact that ENT is more robust with respect to estimation errors. We take into consideration the estimation errors as well as the variations in the channel through the variance term of ENT.

5. Conclusions

This paper developed a new quality-aware routing metric for ad hoc wireless networks, called ENT (for Effective Number of Transmissions). ENT takes into account both the mean and variance of the channel conditions, and is designed to reduce the errors visible to higher layers like TCP after the link layer has attempted to deliver the packet using retransmissions. In fact, ENT is a general concept, and many QAR metrics fit into a unified framework given by our ENT analysis. Network designers can choose from a number of different metrics using the framework we developed. We also showed how nodes in the network can efficiently estimate channel conditions to feed into the network layer.

6. References

[1] Perkins C. E. and Bragwat P., "Destination Sequenced Distance Vector Routing," in *Proceedings of SIGCOMM*, 1994.

[2] Johnson D. B., "Routing in Ad-hoc Networks of Mobile Hosts," in *Proc. of the IEEE Workshop on Mobile Computing Systems and Applications*, 1994.

[3] Bicket J. DeCouto D. S. J., Aguayo D. and Morris R., "A High Throughput Path Metric for Multi-Hop Wireless Routing," in *Proceedings of MobiCom*, 2003.

[4] Park V. D. and Corson M. S., "A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks," in *Proc. of the IEEE Infocom*, 1997.

[5] Yarvis M. et al., "Real World Experiences with an Interactive Ad-hoc Sensor Network," in *Proceedings of the International Workshop on Ad Hoc Networking*, 2002.

[6] Shepard T., "A Channel Access Scheme for Large Dense Packet Radio Networks," in *Proceedings of SIGCOMM*, 1996.

[7] Adya A. et al., "Radio Unification Protocol for IEEE 802.11 Wireless Networks," *Technical Report, MSR-TR-2003-44, Microsoft Research*, 2003.

[8] Willig A., Kubisch M., Hoene Christian, and Wolisz A., "Measurements of a Wireless Link in an Industrial Environment Using an 802.11-Compilant Physical Layer," *IEEE Transactions on Industrial Electronics*, 2003.

[9] Woo A., Tong T., and Culler D., "Taming the Underlying Challenges of Reliable Multihop Routing in Sensor Networks," in *Proceedings of SenSys*, 2003.

[10] Zhao J. and Govindan R., "Understanding Packet Delivery Performance in Dense Wireless Sensor Networks," in *Proceedings of SenSys*, 2003.

[11] Gilbert E. N., "Capacity of a Burst Noise Channel," *Bell Systems Technical Journal*, vol. 39, pp. 1253–1265, 1960.

[12] Elliott E. O., "Estimates of Error Rates for Burst Codes on Burst Noise Channel," *Bell Systems Technical Journal*, vol. 42, pp. 1977–1997, 1963.

[13] Fritchman B. D., "A Binary Channel Characterization Using Partitioned Markov Chains," *IEEE Transactions on Information Theory*, vol. 13, pp. 221–227, 1967.

[14] Willig A., Kubisch M., and Wolisz A., "Measurements and Stochastic Modeling of a Wireless Link in an Industrial Environment," *TKN Technical Report Series, TKN-01-001, Technical University, Berlin*, 2000.

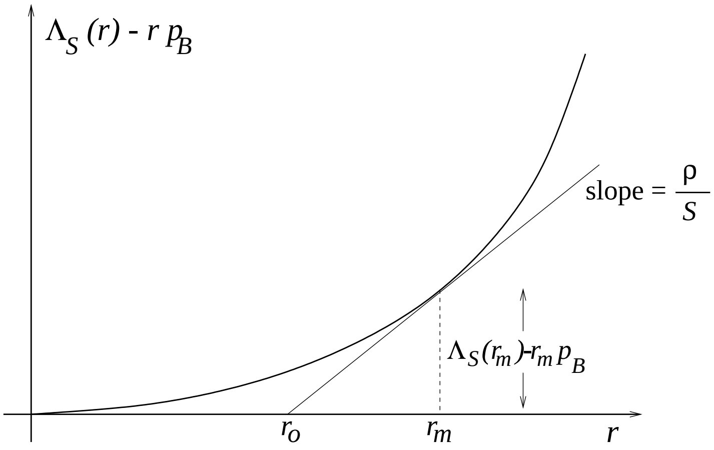

Figure 23: Due to the convexity of $\Lambda_S(r) - r\overline{p}_B$, r_m decreases with decreasing ρ/S.

[15] Woo A. and Culler D., "Evaluation of Efficient Link Reliability Estimators for Low-power Wireless Networks," *Technical Report, UCB//CSD-03-1270, University of California, Berkeley*, 2003.

[16] Kim M. and Noble B., "Mobile Network Estimation," in *Proceedings of ACM MOBICOM*, 2001.

[17] Willsky A., Wornell G. W., and Shapiro J., *Stochastic Processes, Detection and Estimation, 6.432 Course Notes*, Department of EECS, MIT, Cambridge, MA, 1996.

APPENDIX

A. Derivation of the Probability of Packet Loss

In this section, we derive the probability of packet loss, given in (4). Let $\overline{p}_B = \mathrm{E}\left[p_{B,t}\right]$. Then,

$$\begin{aligned}
\log \mathrm{P}\left(\frac{1}{p_c(T)} \geqslant M\right) &= \log \mathrm{P}\left(\sum_{t=T-S+1}^{T} p_{B,t} \geqslant \log M\right) \\
&= \log \mathrm{P}\left(\sum_{t=T-S+1}^{T} p_{B,t} - S\overline{p}_B \geqslant \rho\right) \\
&\leqslant -S\Lambda_S^*\left(\frac{\rho}{S}\right) \qquad (15)
\end{aligned}$$

where

$$\Lambda_S^*\left(\frac{\rho}{S}\right) = \sup_{r\geqslant 0}\left[\frac{\rho}{S}r - \left(\Lambda_S(r) - r\overline{p}_B\right)\right] \qquad (16)$$

and

$$\Lambda_S(r) = \frac{1}{S}\log \mathrm{E}\left[\exp\left(r\sum_{t=1}^{S} p_{B,t}\right)\right]$$

is assumed to be differentiable for $r \geqslant 0$. Inequality (15) is the Chernoff's bound. It becomes tighter as ρ grows. The optimization procedure in Eq. (16) is illustrated graphically in Fig. (23). The abcissa of the point where $\Lambda_S(r) - r\overline{p}_B$ has a derivative of ρ/S is r_m. The tangent line at r_m cuts the x-axis at r_o. Thus,

$$\begin{aligned} r_o &= \frac{S}{\rho}\Lambda_S^*\left(\frac{\rho}{S}\right) \\ &= \frac{S}{\rho}\left[\frac{\rho}{S}r_m - (\Lambda_S(r_m) - r_m\overline{p}_B)\right] \end{aligned}$$

which means,

$$\mathrm{P}\left(\sum_{t=T-S+1}^{T} p_{B,t} \geqslant \log M\right) \approx \exp(-r_o\rho) \tag{17}$$

Recall that $\Sigma = \sum_{t=T-S+1}^{T} p_{B,t}$. From Taylor series expansion, for small values of r,

$$\Lambda_S(r) \approx r\overline{p}_B + \frac{1}{2S}\sigma_\Sigma^2 r^2 \tag{18}$$

since $\Lambda_S'(0) = \overline{p}_B$ and $\Lambda_S''(0) = \sigma_\Sigma^2$. Suppose $r_m \ll 1$. If we use (18) to evaluate r_m we get,

$$r_m = \frac{\rho}{\sigma_\Sigma^2}$$

and

$$r_o = \frac{1}{2}r_m$$

Note that generally, σ_Σ^2 due to the small decay in the autocovariance of $p_{B,t}$ even in very long time scales. If $S\overline{p}_B$ is small, then the Chernoff bound can be very tight even for moderately large values of ρ. Therefore, $r_m \ll 1$ is not an unrealistic assumption and using (18) is valid for $r \approx r_m$. Rewriting Eq. (17),

$$\mathrm{P}\left(\sum_{t=T-S+1}^{T} p_{B,t} \geqslant \log M\right) \approx \exp\left(-\frac{1}{2}\frac{\rho}{\sigma_\Sigma^2}\cdot\rho\right) \tag{19}$$

B. Combining ENT Estimates

Given the estimates, $\hat{\alpha}(\delta, \tau + kS)$, of the ENT for individual time periods, $0 \leqslant k \leqslant T_r/S$, we show how to come up with a single metric to find the best route for the entire routing period.

Claim 1. *If $\hat{\Sigma}_k$ is an unbiased estimate of Σ_k and the corresponding mean squared error is ϵ_k, then mean squared error corresponding to $\hat{\Sigma}_k + A$ is $\epsilon_k + A^2$.*

Proof: Let $\hat{\Sigma}_k$ be an unbiased estimate. Then,

$$\begin{aligned} \mathrm{E}\left[\left(\Sigma_k - (\hat{\Sigma}_k + A)\right)^2\right] &= \mathrm{E}\left[(\Sigma_k - \hat{\Sigma}_k)^2\right] \\ &+ 2A\mathrm{E}\left[\Sigma_k - \hat{\Sigma}_k\right] + A^2 \\ &= \epsilon_k + A^2 \end{aligned}$$

completing the proof. Thus, if we use $\hat{\Sigma}$ instead of $\hat{\Sigma}_k$ for $0 \leqslant k \leqslant T_r/S$, the total mean squared error is incresed by $\sum_{k=1}^{T_r/S}(\hat{\Sigma} - \hat{\Sigma}_k)^2$. The choice of Σ that minimizes this amount of increase is

$$\hat{\Sigma} = \frac{1}{T_r/S}\sum_{k=1}^{T_r/S}\hat{\Sigma}_k$$

Hence, the "best" log ENT estimate for the period $(\tau, \tau + T_r)$ is

$$\hat{\alpha}(\delta) = \frac{1}{T_r/S}\sum_{k=1}^{T_r/S}\left[\hat{\alpha}(\delta, \tau + kS) + 2\delta(\hat{\Sigma} - \hat{\Sigma}_k)^2\right]$$

C. EWMA Filters

In this section we discuss the EWMA for estimating ENT. EWMA estimator is the main tool used in the context of loss rate estimation. In general the EWMA is used in the following form (EWMA for $1_{B,t}$ is a trivial extension of this):

$$\hat{\Sigma}_k^{\mathrm{ewma}} = S\sum_{s=0}^{\tau}(1-\gamma)\gamma^{\tau-s}p_{B,s} \tag{20}$$

where $0 < \gamma \leqslant 1$. Moving averaging is the degenerate case of EWMA where $\gamma = 1$. We call this the the generic EWMA. Let $\vec{\Gamma} = [\gamma^{\tau-1}\ \gamma^{\tau-2}\ \cdots\ 1]'$. The error variance associated with the generic EWMA is,

$$\begin{aligned} \epsilon_k^{\mathrm{ewma}} = \sigma_{\Sigma_k}^2 &- \left\{2S(1-\gamma)\left(\vec{\Gamma}\cdot\vec{K}_{\Sigma_k\vec{p}_B}\right)\right. \\ &- \left.\mathrm{var}\left(S(1-\gamma)\left(\vec{\Gamma}\cdot\vec{p}_B\right)\right)\right\} \end{aligned} \tag{21}$$

where "·" represents inner product. The generic EWMA does not take advantage of the known first two order statistics. In fact, EWMA estimate is unbiased, that is, it inherently performs mean estimation. It also assumes decaying autocovariance by putting more weight on the recent observations. Thus, it is designed for the cases where statistics are not known or are not estimated accurately. Later in this paper we will analyze the generic EWMA and show that its performs poorly (worse than MA) in real world scenarios and toy examples.

We propose a modified version of EWMA which uses the first two order statistics.

$$\hat{\Sigma}_k^{\mathrm{ewma}} = S\left[\bar{p}_B + \frac{\|\vec{K}_{\Sigma_k\vec{p}_B}\|}{\sigma_\Sigma^2}\sum_{s=0}^{\tau}(1-\gamma)\gamma^{\tau-s}\left(p_{B,s} - \bar{p}_B\right)\right] \tag{22}$$

The error variance can be evaluated as:

$$\begin{aligned} \epsilon_k^{\mathrm{ewma}} &= \sigma_{\Sigma_k}^2 - \left\{2\frac{\|\vec{K}_{\Sigma_k\vec{p}_B}\|}{\sigma_\Sigma^2}(1-\gamma)\left(\vec{\Gamma}\cdot\vec{K}_{\Sigma_k\vec{p}_B}\right)\right. \\ &- \left.\mathrm{var}\left(\frac{\|\vec{K}_{\Sigma_k\vec{p}_B}\|}{\sigma_\Sigma^2}(1-\gamma)\left(\vec{\Gamma}\cdot\vec{p}_B\right)\right)\right\} \end{aligned} \tag{23}$$

It is clear that given the first two order statistics, one could use the LLS estimator instead of the modified EWMA. However, the matrix inversion in the LLS estimator may be undesirable especially if the probing interval is large.