Link Quality Estimation for Data-Intensive Sensor Network Applications

Parya Moinzadeh, Kirill Mechitov and Gul Agha
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Avenue, Urbana, IL 61801
{moinzad1, mechitov, agha}@illinois.edu

Abstract

The efficiency of multi-hop communication is a function of the time required for data transfer, or throughput. A key determinant of throughput is the reliability of packet transmission, as measured by the packet reception rate. We follow a data-driven statistical approach to dynamically determine a link quality estimate (LQE), which provides a good predictor of packet reception rates. Our goal is to enable efficient multi-hop communication for applications characterized by data-intensive, bursty communication in large sensor networks. Statistical analysis and experiments carried out on a network of 20 Imote2 sensors under a variety of environmental conditions show that the metric is a superior predictor of throughput for bursty data transfer workloads.

1. Introduction

As the scale of the sensing systems grows and sophisticated applications demand more and better data, the communication capabilities of low-power embedded sensors are pushed to the limit. Communication issues emerge in these data-intensive applications which do not manifest in WSN systems such as environmental monitoring and target tracking which involving sporadic or low-frequency collection of small amounts of data [9], [6], [8].

Multi-hop communication is necessary in WSNs where the network covers an area too large to allow direct communication between all sensors. Routing protocols try to find the best sequence of intermediate nodes to successively forward the packets to the destination—using metrics such as hop count or total energy consumption—and a data transport protocol then uses these routes to send data. The performance of high-throughput, bursty data transfers in multi-hop networks is particularly sensitive to the quality of the routes generated by the routing protocol, as packets dropped on each hop have a multiplicative effect on the overall packet loss rate.

In some applications, distributed sensors collect non-redundant data and must employ a reliable data transport protocol to avoid losing significant information due to packet loss. Damage detection in structural health monitoring is one such application, where packet loss may induce false detections. Packet acknowledgments and resending of dropped packets are typically used to recover from packet losses. In these applications, communication performance also plays a significant role in energy consumption, as lower throughput requires the nodes in the network to consume energy not only to resend the dropped packets, but also to keep the node running for the longer duration required for the data transfer.

Link quality has been identified as an important determinant of the multi-hop routing algorithm performance in sensor networks [17], [4], [12], [14]. Link quality varies not only with transmission power and the distance between nodes, but with environmental conditions which are likely to be different from one WSN deployment to another. Thus a single formula or heuristic may not be sufficient to estimate link quality with accuracy.

This work identifies critical factors affecting the performance of multi-hop routing in data-intensive WSN applications. We use a data-driven statistical model to calculate a low-cost routing metric based on hardware link quality indicators of IEEE 802.15.4 radios commonly used in WSN platforms. The novelty of our method is that it combines information from these indicators using a regression tree model. Our method significantly reduces the prediction error when compared to the construction of a singular model for the entire data space (as is common in the literature). When integrated with a reliable data transport protocol tailored to the unique requirements of data-
intensive sensor network applications (e.g. [10]), this metric provides higher quality, more stable routes with significantly improved performance measures such as packet loss rate and throughput.

The remainder of the paper is organized as follows. Section 2 surveys the related work on link quality estimation in WSNs. In Section 3, we motivate this research by examining the unique challenges posed by high-throughput sensor network applications. Sections 4 and 5 present our analysis of link quality and an original algorithm for calculating the routing metric. Results of experimental evaluation on a network of Imote2s are presented in Section 6, along with a discussion of their impact on routing and data transport protocols. Section 7 concludes the paper.

2. Related Work

Several multi-hop routing algorithms have been proposed specifically for sensor networks, most commonly distinguished from those for ad hoc wireless networks by when and how route information is generated and how it is updated [13], [18], [1], [3]. The choice of routing metric plays an important role in these protocols. An efficient routing metric should address energy limitations, link quality variation, and diverse radio environments.

In [4], the authors design a routing metric, called RLQ, which is based on both energy efficiency and link quality statistics. It aims to adapt to link quality variations, and node heterogeneity in wireless sensor and actor networks. A stable link quality metric for WSNs is developed in [12]. Based on experimental evaluations on the behavior of different links, the authors present an indicator that combines received signal strength and packet reception rate. However, the metric does not account for signals that are near the radio’s sensitivity threshold, as in such areas there is no consistent relationship between the two values. In [17], link qualities between nodes are estimated by measuring the received signal strength of the control messages, and a path with minimum sum of the values is chosen. As mentioned by the authors, the proposed method may not minimize energy consumption. This method also does not account for short-term link instability as noted in [14].

A bidirectional sink-based routing protocol is presented in [16], with ETX (expected transmission count) as the chosen routing metric. ETX is commonly used in 802.11 wireless networks [2]; however, this solution is not optimal for networks operating on 802.15.4 radios, as they are prone to significant and instantaneous link quality changes [14].

3. Motivation

In order to understand the requirements of multi-hop communication, we classify communication in sensor networks into two categories:

1) *Low throughput communication*. Example applications include medical monitoring, environmental observation and forecasting systems, and habitat monitoring [9], [6], [8]. These applications require relatively low-frequency data gathering and experience large variations in radio communication and node locations.

2) *High-throughput and bursty data transfer communication*. Example applications include monitoring the structural health of civil infrastructure and rare event detection (earthquake, mudslide, etc.) [5], [11]. Applications in this category impose specific requirements such as high sampling rates, timely data collection and analysis, a large volume of data, precise inter-nodal synchronization, and reliable communication.

In this paper, we focus on high-throughput applications with bursty data transfer requirements; our goal is to understand how to estimate link quality. Because link quality indicates the likelihood of successful packet delivery across the link, link quality estimation enables the creation of highly reliable and energy-efficient routes.

We now examine the features that distinguish data-intensive WSN applications from the networking standpoint. Sensor networks are known for challenging resource requirements, including limited radio bandwidth, computational power, and energy. Nevertheless, data-intensive applications have proven to be necessary in several application domains. This in turn makes performance—efficiency in the exploitation of
the resources—of paramount concern. As we look at networking performance, the following concerns must be taken into account in the design of routing and data transfer protocols for data-intensive applications.

**Bursty communication.** Communication in most WSN applications involves exchanging commands and limited amounts of data sporadically or at low frequency, both of which are typically small enough to be encapsulated in single packets. However, data-intensive applications often need to transfer large amounts of measurement data among sensor nodes in a limited amount of time. Combined with high sampling rates, this usually precludes real-time data collection, as even a small number of nodes can easily saturate the available network bandwidth. As a result, sensing and data collection take place sequentially, and communication is bursty—periods of intense communication are interspersed with virtually no communication activity.

**Reliability.** Reliability of transporting acquired sensor data is vital in many applications. Most scenarios assume measurement data is available from all the nodes without intermittent loss. Packet loss compensation is therefore required [11].

**Heterogeneous link quality.** The radio communication environment at the location where the sensor network is deployed can be complex due to RF reflection, refraction, absorption, and other phenomena. Structures such as buildings and bridges consist of numerous components made from steel, concrete or other materials that have varying and often unpredictable effects on the maximum range of communication and the strength of the radio signal. In outdoor environments such as fields or forests, moisture content in the vegetation likewise affects radio wave propagation. Therefore, the communication range and link quality will vary from place to place, and its estimation prior to on-site testing is challenging.

Figure 1 shows the direct relationship between link qualities, the nodes were placed at different locations, and poor-quality links are prevalent. Link reliability when the network topology is sparse, ratio. It is expected to have a higher correlation with the received energy level and/or the signal-to-noise ratio. LQI is a composite value intended to characterize link quality, measuring the received energy level and/or the signal-to-noise ratio. It is expected to have a higher correlation with link reliability when the network topology is sparse, and poor-quality links are prevalent.

Let us investigate the relationship between LQI and RSSI, as well as RSSI and PRR. Table 1 summarizes the details of our experimental setup. Our experiments were performed in three different conditions and represent different radio communication environments. For each experiment, in order to gather data for different link qualities, the nodes were placed at different locations with respect to a transmitting node. Figure 2 shows the results for the Office dataset. These results

**4. Link Quality Estimation**

First, we summarize the building blocks of our link quality estimator design. Following this overview, we will discuss the details of link quality estimation and metric calculation.

**Passive link estimation.** The requirement of minimal network overhead identified above strongly implies that the link estimation algorithm should not rely on periodic update messages. We refer to such link quality estimation techniques as passive. There are two major reasons behind this design. First, sending periodic update messages imposes a significant energy consumption cost. In order for the cost of periodic link quality assessment to be small compared to that of data transfer, the interval between updates should be long. However, infrequent updates are based on the assumption of link stability, which does not hold for 802.15.4 [14]. Second, during high-throughput data transfers the frequent update messages interfere with the transport protocol traffic.

**Packet reception rate as metric.** Our primary focus is on high throughput, bursty communication. Since there is a direct relationship between packet reception rate (PRR) and throughput (see Figure 1), predicting the link PRR will in turn determine the maximum throughput that the link can support.

**Agile link assessment.** Link instability and time-variance of radio conditions suggest that rapid assessment of link quality based solely on short-term information is critical. Our protocol makes that assessment based on physical layer parameters measured over a single packet, namely RSSI (Received Signal Strength Indicator) and LQI (Link Quality Indicator). RSSI is the estimate of the signal power and is highly correlated with packet reception rate, except when operating at the edge of receiver sensitivity. LQI is a composite value intended to characterize link quality, measuring the received energy level and/or the signal-to-noise ratio. It is expected to have a higher correlation with link reliability when the network topology is sparse, and poor-quality links are prevalent.

Let us investigate the relationship between LQI and PRR, as well as RSSI and PRR. Table 1 summarizes the details of our experimental setup. Our experiments were performed in three different conditions and represent different radio communication environments. For each experiment, in order to gather data for different link qualities, the nodes were placed at different locations with respect to a transmitting node. Figure 2 shows the results for the Office dataset. These results
are in agreement with previous work in [15] and [7] for MicaZ, [4] for Tmote Sky nodes, [12] for Moteiv’s TelosB, and [14] for Telos revB and MicaZ. In this paper, we do not aim to provide a complete analysis of the low power wireless behavior and rather focus on its implications on link estimation and routing. According to these graphs and previous results on 802.15.4, RSSI has high correlation with PRR when it is above -87 dBm, and shows little correlation at the edge of radio sensitivity.

As an aggregate metric, LQI averaged over many packets is strongly correlated with the packet reception rate; however, due to high variance, individual LQI readings are not sufficient to confidently estimate the PRR. Figure 3 shows LQI scaled to the range of 0–100 (to estimate packet reception rate) versus the measured PRR values for the same data sample. To estimate the PRR based on LQI, plot a horizontal line at a given LQI value, then plot a vertical line at the intersection with the LQI curve. The intersection of this line with the PRR curve gives the actual PRR for this LQI estimate. As seen in the figure, however, in many cases there are multiple such intersections, with the measured PRR value varying by as much as 40 percentage points.

To illustrate the significance of LQI and RSSI variables for PRR estimation, we first ran a linear regression with PRR as response variable, and LQI as the regress variable on our Office data set. The coefficient of correlation for LQI was equal to 0.9954, showing that the variation in PRR is reduced by 99.54 percent when LQI is considered. However, the average squared error was 119.47, which is quite high. We then ran another regression on the resulted error from the regression on LQI to assess the effect of RSSI on PRR prediction. The coefficient of correlation for RSSI was 0.9579, showing the significance of RSSI in reducing PRR prediction error.
5. Link Metric Computation

We aim to develop an accurate model for predicting PRR using RSSI and LQI variables. First-order regression is an attractive solution due to its simplicity and low computation cost.

We first consider a global multiple linear regression model with two regressor variables RSSI and LQI, and dependent variable PRR. The linear regression model and the transformation for regressor variables are given by:

\[ Y = X\beta + \epsilon \]  
\[ \log(\mu/(1\mu)) = Xb \]

where \( X \) is the input matrix containing experimental RSSI and LQI values, \( Y \) is a vector of predictions for PRR, and \( \epsilon \) is a vector of error terms.

In order to measure the actual predictive capacity of the selected regression models, we use the resulting model to predict each case and then calculate the mean of the squared prediction errors, denoted by \( MSPR \). \( MSPR \) is determined by the following formula:

\[ MSPR = \frac{\sum_{i=1}^{n\ast}(Y_i - \hat{Y}_i)^2}{n\ast} \]

where \( Y_i \) is the value of the response variable in the \( i \)th validation case, \( \hat{Y}_i \) is the predicted value for the \( i \)th validation case based on the model-building data set, and \( n\ast \) is the number of cases in the validation data set.

Table 2 shows the error for PRR prediction via linear regression. The overall prediction error using multiple linear regression is relatively low. However, as shown in Table 2, the error for communication on the edge of radio sensitivity is very large. This behavior is undesirable for communication on the edge of radio sensitivity and is expected. The large and inconsistent error observed in prediction results was expected due to the different distribution of RSSI and LQI data for different link qualities. Moreover, results from different datasets encounter large variations in the error. This inconsistency can be explained by the nature of these different datasets. The Bridge dataset was gathered in an open area with direct line of sight between most node pairs and little to no radio interference, and thus contains fewer data points for links that are at the edge of radio sensitivity. Since the inaccuracy of multiple linear regression method mostly arises in this portion of the data space, observing smaller errors for datasets that mostly contain data from good communication environments is to be expected.

When data has multiple features with complicated and nonlinear interactions, assembling a single global prediction model can be very difficult and error-prone. An alternative approach to a single regression model is to recursively partition the space into smaller regions, until we reach sub-regions with approximately linear behavior. We are specifically interested in deriving multiple linear models as their computation imposes very little overhead, and are not likely to incur overfitting. It is important to note that here we do not aim to maximize the correlation between attribute values in each subspace, but rather seek to maximize the prediction accuracy of the response variable (PRR) via data space partitioning. This recursive partitioning of the data space can be envisioned through regression trees, which are similar to binary decision trees. Each inner node of the regression tree represents a question about an attribute value, based on which the data is partitioned. Tree edges are labeled with yes/no answers to the questions, and the leaf nodes are labeled with values representing each class. However, classic regression trees only provide a constant estimate of the predicted value on each leaf node. This approach has the advantage of simplicity and ease of implementation but may lead to increased prediction error. We perform a small modification to the regression tree, in that we add linear models on the leaf nodes.

Figure shows 4 part of the regression tree built for the Office dataset. The complete tree consists of 40 levels and correctly represents the interaction between RSSI, LQI and PRR. Once the tree is fixed, the local models can be easily determined. Thus, the major ef-

---

**Algorithm 1** Regression tree algorithm for PRR estimation

/* Off-line model building */
\[ t = \text{RegressionTree}(\text{RSSI}, \text{LQI}, \text{PRR}) \]

/* Prune at level \( L \), with error bound \( \alpha \) */
\[ t_{\text{pruned}} = \text{Prune}(t, L, \alpha) \]

for all \( d_i \in \text{Leaves}(t_{\text{pruned}}) \) do
  if \( \text{Size}(d_i) \geq \beta \) then
    \[ m_i = \text{Regres}(d_i) \]
  else
    \[ m_i = \text{Average}(d_i) \]
  end if
end for

/* On-line PRR prediction for packet \( p \) */
\[ i = \text{Search}(t_{\text{pruned}}, \text{LQI}_p, \text{RSSI}_p) \]
\[ \text{PRR}_{\text{link}}(p) = m_i(\text{LQI}_p, \text{RSSI}_p) \]
Table 2. Error of multiple linear regression for PRR prediction using LQI and RSSI

<table>
<thead>
<tr>
<th>Dataset</th>
<th>All regions</th>
<th>RSSI &gt; -85dBm</th>
<th>RSSI &lt;= -85dBm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSRP</td>
<td>Max Error</td>
<td>MSRP</td>
</tr>
<tr>
<td>Office</td>
<td>202</td>
<td>49</td>
<td>43</td>
</tr>
<tr>
<td>Bridge</td>
<td>30</td>
<td>42</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 3. Error of regression tree for PRR prediction using LQI and RSSI

<table>
<thead>
<tr>
<th>Dataset</th>
<th>All regions</th>
<th>RSSI &gt; -85dBm</th>
<th>RSSI &lt;= -85dBm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSRP</td>
<td>Max Error</td>
<td>MSRP</td>
</tr>
<tr>
<td>Office</td>
<td>45</td>
<td>59</td>
<td>8</td>
</tr>
<tr>
<td>Bridge</td>
<td>27</td>
<td>47</td>
<td>12</td>
</tr>
</tbody>
</table>

Figure 4. Part of the regression tree built for Corridor dataset. The inner nodes partition the data space based on two attributes: RSSI and LQI. Each leaf represents a subspace of the entire dataset and is associated with a value representing the average response variable (PRR) for data points in that subspace.

6. Evaluation

In this section we present our results on link quality estimation. For this purpose, we apply models generated during the training phase to new experimental data, which is gathered in the same environment as the training data. Our purpose is to assess the applicability of the derived link quality estimation models for the sensor network deployment lifetime.

Table 5 shows the characteristics of the generated regression tree for the Office data set. To optimize metric computation, the number of regression tree levels is reduced by pruning the tree while keeping the error at a fixed bound. Figure 5 compares PRR estimation via a unified multiple linear regression model, and the regression tree. While the error is significantly reduced by the tree structure, it is still high for poor radio environments. The relatively high error occurs in subspaces that contain a small number of samples, since the regression model can not provide a valid estimate. To
tackle this problem, we limit the use of the regression model to subspaces that contain a sufficient number of samples. This restriction can also be applied by allowing regression modeling only when the standard error for the regression is less than a threshold $\beta$. For the subspaces where regression is not performed, a simple average is used as the predictor. Figure 5 shows the resulting regression tree model. From the figure we can see that the modified regression tree model is a significant improvement over the original tree. In order to investigate the effect of data space partitioning and using both RSSI and LQI as regress variable we have compared our model with models that consider raw LQI [4], and raw RSSI values [15]. Figure 6 shows these results. Among these models, LQI gives the best results when used as the regress variable. Compared to regression based on raw LQI values, the regression tree model reduces the PRR estimation error by 67%, on average. The large variance of the estimation in bad communication environments is due to the high noise variations on different nodes. Table 4 summarizes these results.

To better assess the applicability of our link estimation model for multi-hop routing in WSN applications, we use our model to categorize link qualities based on PRR. Figure 7 represents our definition of link quality classification which is based on our application requirements and the relationship between PRR and throughput. This classification scheme allows the rout-
ing algorithm to use criteria besides link quality (e.g., shortest path or minimum energy) within the categories without significantly affecting the reliability of routes. In this classification scheme, significant errors, where a link was classified neither to its exact group or one of its neighboring groups, appeared in only 4 out of 266 cases. This confirms that the regression tree method provides an accurate classification of link quality, suitable for use in routing and data transport algorithms.

7. Conclusions and Future Work

Data-intensive WSN applications that feature high-throughput, bursty communication present a challenging environment for multi-hop wireless communication. The link quality metric we have developed enables routing and data transport to select routes best suited to this task. Automated, data-driven regression tree method to derive the link estimation metric can aid routing and data transport protocols operating under to changing communication conditions.

Aspects of the regression tree approach can be further refined in future work. Rather than falling back to using the category average when there is insufficient data to perform regression, interpolation, filtering and smoothing can be applied. Furthermore, tuning the parameters of the regression tree at run-time would permit the metric to adjust more rapidly to time-varying environment changes.

8. Acknowledgments

The authors gratefully acknowledge the support of this research by the National Science Foundation, under grant CMS 0600433 and CNS 1035773.

References


