Comparison of Parametric and Non-Parametric Statistical Features for Z-Wave Fingerprinting

Hiren J. Patel  
Cyber Integration and Transition Branch  
Air Force Research Laboratory  
Rome, NY, 13441  
Email: hiren.patel@us.af.mil

Benjamin W. Ramsey  
Department of Electrical and Computer Engineering  
Air Force Institute of Technology  
Dayton, OH, 45433  
Email: benjamin.ramsey@afit.edu

Abstract—The number of internet connected devices by all accounts is set to increase dramatically in coming years as Internet of Things technologies become cheaper and more convenient. Z-Wave devices have found application in building control, smart energy, health care and equipment monitoring. Its closed standard ensures interoperability of devices and this stability has led to its popularity among consumers. As use of these devices becomes more widespread, the need to protect them becomes more important. In this research, the RF-DNA fingerprinting method is examined to protect these devices using their physical layer attributes. In particular, the traditional method of using parametric features such as variance, skewness, and kurtosis is challenged with the use of non-parametric features mean, median, mode and linear regression coefficient estimates. With careful analysis of variables, a 71% reduction in features is achieved while attaining >94% correct classification rate at an 8 dB lower SNR than using traditional parametric features.  

I. INTRODUCTION

Z-Wave is an inexpensive wireless solution with growing popularity in security systems, building automation, smart energy, and other networked applications. However, the proprietary and closed source nature of the Z-Wave protocol makes security evaluations challenging. Proof-of-concept attacks against Z-Wave networks have been demonstrated using customized transmitters built from software-defined radios and sub-GHz development kits. In light of these attacks, wireless situational awareness is an important component for network defense-in-depth. Radio frequency (RF) fingerprinting with low-cost signal collection receivers to protect critical infrastructures was recently demonstrated for the low-rate ZigBee protocols [1], but this work is the first to examine such techniques on Z-Wave.

RF fingerprinting offers a unique method of authenticating devices using physical layer security. Minute differences in the physical structure of these devices such as in the oscillator, amplifier, and delay circuits manifest into differences in the amplitude, phase and frequency of a transceiver. While these differences are minute and still produce adequate symbol constellations for digital communication, they have been found useful for device authentication [2]–[5]. The RF signal from each network node is collected, reduced with feature generation and used to generate a device authentication model via various classifiers. With the learned model, these devices can then be identified and authenticated. Many of the methods in this area of research use parametric feature generation methods to summarize the collected RF signal from network nodes. This works demonstrates that parametric feature generation methods frequently result in inferior classification when compared to other non-parametric methods. Non-parametric feature generation methods are tested herein with an established RF-fingerprinting process to quantify device authentication accuracy.

The remainder of this paper is organized as follows: Section II provides background information on RndF and feature generation. Section III describes the methodology for data collection, and device authentication. Section IV provides experimental results and analysis for experiments on device authentication. Finally, conclusions from this work are detailed in Section V.

II. BACKGROUND

A. Random Forest Entropy Importance

The RndF algorithm is used herein as it is empirically demonstrated to be robust for high-dimensional multi-class problems [6]. RndF has a unique variable importance metric built in. The RndF consists of a number of decision trees, each of which is made up of a number of decision nodes. During model development in the training phase, a handful of variables are randomly chosen at each node $a$. Then, an impurity metric is calculated from Shannon’s Entropy for $c$ classes, given by

$$ I(a) = -\sum_{j=1}^{c} p_a(j) \log_2 p_a(j) $$  \hspace{1cm} (1)$$

where the class probability $p_a(j)$ is efficiently calculated by dividing the number of observations in class $a$ over the total number of observations at that node. For each variable of these observations, all possible threshold values are examined. For a variable $x_i$ and threshold $t_i$, the training set of observations at the parent node is split into two sets of child nodes, with observations where $x_i \geq t_i$ going to one child node, and the remaining going to the other child node. Then the reduction
in impurity between the parent and child nodes is noted. The variable and threshold combination \( \{ x^*, t^* \} \) that yields the best impurity reduction is kept. Thus RndF chooses the best variables from a subset for training a decision node. The impurity reduction can be averaged for all nodes in all trees and a relative metric of variable importance is generated. More information can be found in [7], [8]. Here the RndF variable importance metric pre-selects variables that provide the most information gain in terms of entropy reduction.

**B. RF Fingerprinting**

Steady-state RF-fingerprinting methods have generally used the preamble portion of the transmitted data packet for feature generation [3], [9]–[16]. In [3], [15], [16] a method called RF-DNA fingerprinting is used where the input signal is divided into several equal size Regions of Interest (ROI). In each ROI, statistics of variance, standard deviation, skewness, and kurtosis are calculated and these serve as input features to the classifier for device authentication. However, inaccuracies exist in this method when these Gaussian distribution statistics are calculated over ROIs exhibiting clearly non-Gaussian distributions, as shown in Figure 1. Here, the 80 bit preamble is divided into 40 equal-size ROIs. Each ROI includes two bits of the preamble sequence. All ROIs are investigated and found to exhibit similar non-Gaussian behavior. Non-parametric feature generation methods are hypothesized to perform better if they can more accurately summarize the ROIs.

In [9], the Power Spectral Density (PSD) coefficients of the preamble are proposed as a non-parametric method for features generation. In [12]–[14] Fourier coefficients are used for classification. These methods require sufficient number of coefficients to provide proper differentiation between signal and noise frequencies. Often, due to the high center frequencies of wireless transmissions, a large pass-band frequency range is used by the receiver. These situations require the generation of large numbers of PSD or Fourier coefficients which can be difficult for many classifiers to use.

There is a need for a robust non-parametric feature generation method that accurately summarizes the received steady-state preamble signal. Current frequency-based methods using PSD or Fourier coefficients do not provide a closed loop solution in terms of identifying how many coefficients are sufficient for classification. Therefore, in this work other non-parametric statistics are investigated, to include: mean, median, and mode of the ROIs. In addition, although linear regression assumes a normal noise distribution around the model estimate, the trend of non-Gaussian data can still be accurately estimated via a linear regression model. In this case, the coefficient estimates of the linear model can serve as features to the classifier. RF-fingerprinting with these linear regression coefficients as features is also investigated.

III. SIGNAL COLLECTION METHODOLOGY

Three Z-Stick S2 transmitters are the devices under test in this research. An NI USRP-2921 software-defined radio serves as a low-cost signal collection receiver. In-phase and quadrature (I/Q) samples stream from the USRP as 16-bit integers, sampled at 2 Msps, and are stored on a Dell Precision M4500 laptop running LabVIEW. The I/Q binary file format is an interleaved array

\[
[I_0, Q_0, I_1, Q_1, \ldots, I_n, Q_n],
\]

where \( n \) is the number of collected I/Q sample pairs. Interleaved I/Q data is converted to complex values in the format

\[
[I_0 + Q_0, I_1 + Q_1, \ldots, I_n + Q_n],
\]

for signal processing in MATLAB. A total of 230 Z-Wave preambles (9.6 Kbps data rate, one transmission per second) are collected from three Aeotec Z-Stick S2 transmitters. Transmission detection from background noise is accomplished through amplitude-based leading edge detection using a -6 dB threshold. As outlined in the ITU-T G.9959 specification [17], the first 8.3 ms of each transmission constitutes the preamble; at 2 Msps the first 17667 instantaneous I/Q samples represent the preamble region.

The Z-Stick S2 transmitters are placed 10 cm from a vertically-oriented LP0410 log periodic antenna, connected via cable directly to the RF input of the USRP. Collected signal-to-noise ratio (SNR) is 24 dB. To determine the effect of noise on classification performance, White Gaussian Noise (WGN) is generated to achieve \( \text{SNR} = \text{[0,24]} \) dB at 2 dB intervals. This WGN is filtered with the previously used Butterworth filter prior to adding it to the signal.

Feature generation is conducted by dividing the 80-bit preamble into 40 equal size ROIs, such that each ROI corresponds to two bits. The complex I-Q variables are converted to instantaneous amplitude \( a[n] \), phase \( \phi[n] \) and frequency \( f[n] \) variables by

\[
[\{ I_0, Q_0 \}, \{ I_1, Q_1 \}, \ldots, \{ I_n, Q_n \}],
\]
\[ a[n] = \sqrt{I[n]^2 + Q[n]^2}, \]
\[ \phi[n] = \tan^{-1}\left(\frac{Q[n]}{I[n]}\right), \quad \text{for } I[n] \neq 0, \]
\[ f[n] = \frac{1}{2\pi} \left[ \frac{d\phi[n]}{dt} \right]. \]

Two sets of features are generated for each set of amplitude, phase, and frequency variables: parametric features and non-parametric features. For the parametric features, variance, skewness and kurtosis functions are generated for each ROI and each observation. The functions of mean, median, mode, slope and intercept estimated by linear regression are used as the non-parametric features. Linear regression is performed for each ROI and each observation. As this is a time-based signal, the the independent variables are the time samples \( t_s \) of each collected sample. As samples are collected at equal time intervals, we simply label the x-axis indices of the independent time variable from 1 to \( N_{ROI} \), which is the number of samples in the ROI. Figure 2a shows a plot of an ROI for observations from 3 devices, and the lines fitted through linear regression for each device. This figure gives good insight into how the magnitude of the phase varies per device. The estimated slope appears to be nearly equal for this ROI, however the intercepts are much different per device as they are directly related to the phase magnitude. Indeed, if the means are removed as in Figure 2b we see that the fitted lines for the three devices are very similar. This figure gives a good indication that simply the mean of an ROI per device may be enough to provide reliable features for device authentication.

Classification is performed for each SNR, and for each feature generation methods, as highlighted in Algorithm 1. The Random Forest classifier is used here as it performs very well with non-parametric data and in high dimensional data sets [8]. Each Random Forest consists of 1000 trees, as this size was found to be most useful in prior research [3].

**Algorithm 1 Feature generation for RF-fingerprinting**

Choose subset of \{amplitude, phase, frequency\}
Choose subset of \{variance, skewness, kurtosis, mean, median, mode, linear regression estimates\}

for SNR = 0 to 24 dB in 2 dB steps do

Divide data into train/test sets
Train Random Forest authentication model
Classify test set
Record correct classification rate
end for

IV. RESULTS AND ANALYSIS

A. RndF Variable Importance Results

As previously mentioned, RndF has been empirically shown to perform well in high dimensional data sets. Initially, the parametric feature set with variance, skewness and kurtosis features from each ROI is used to train a RndF. The variable importance for this set is shown in Figure 3. This figure clearly shows that all amplitude features are more important than phase or frequency features toward classification. By using only these, we get a 66% reduction in variables in our training set.

Similarly, the non-parametric features with mean, median, mode, and linear regression slope and intercept estimates are used to train another RndF. The variable importance is shown in Figure 4. Here the amplitude mean and median are more important than other metrics. This enforces the earlier hypothesis in Section III that the Z-Wave devices are distinguishable by their amplitude bias. Frequency linear regression component features are also found to be useful for accurate classification. Finally, the parametric and non-parametric features are concatenated and used to train a third RndF. By training the same classifier with both sets of features, a comparison can be made between the parametric and non-parametric sets. Figure 5 shows that the non-parametric features are given slightly higher preference than the traditional parametric features. A reduced set using the Amplitude variance, skewness and kurtosis features, and Amplitude mean, median and linear regression features is used for further classification testing.
Fig. 5: RndF variable importance for the combined set with parametric and non-parametric features at the collection SNR of 24 dB. Major categories of amplitude, phase and frequency features are labelled.

Fig. 3: RndF variable importance for parametric features at the collection SNR of 24 dB. Major categories of amplitude, phase and frequency features are labelled. Each major category is further divided into sections for variance, skewness and kurtosis.

Fig. 4: RndF variable importance for non-parametric features at the collection SNR of 24 dB. Major categories of amplitude, phase and frequency features are labelled. Each major category is further divided into sections for mean, median, mode, and linear regression slope and intercept components.

B. Device Identification Results

Device identification success is measured by the classification rate. WGN is injected to achieve a variety of SNRs and for each device classification is performed. Training and test sets are equal size and consist of 130 observations per device. Correct classification rates are plotted in Figure 6. A full feature set, all parametric features, all non-parametric features and the reduced set from section IV-A are compared. Results show that the reduced set outperforms or matches the full feature set with 71% fewer variables. At 14 dB 100% correct classification rate is achieved. The full feature set and the non-parametric feature set perform nearly identically, suggesting that the non-parametric features are very useful toward device identification. The traditional parametric only feature set underperforms the remaining at all SNRs.

Table I displays the first 5 SNRs tested and best performance is marked in bold. Under all SNRs, the parametric feature set performs the lowest in terms of correct classification rate. The reduced feature set first passes 90% correct classification at 10 dB at 94%, while the traditionally used parametric feature set does so at 18 dB. Thus this reduced set provides an 8 dB gain by providing similar classification performance at a lower SNR.

V. CONCLUSIONS AND FUTURE WORK

In this work, new non-parametric feature generation methods are investigated with RF-fingerprinting of Z-Wave devices. To our knowledge, this is the first time these devices have been investigated for PHY layer authentication with RndF. It is

Approved for Public Release; Distribution Unlimited: 88ABW-2015-1475
found that there is significant improvement by using features such as mean, median, mode and linear model coefficient estimates of a ROI versus previously used variance, skewness and kurtosis which follow a Gaussian distribution assumption. Careful analysis prior to classification revealed that the signal collected at an ROI does not follow a Gaussian distribution, but that differences in devices are likely due to an signal offset difference. Using the reduced feature set determined with RndF variable importance analysis, an 8 dB gain is achieved for the 90% correct classification rate. This research shows the advantage of signal analysis prior to classification to determine when non-parametric methods can offer better authentication performance.

While these new non-parametric methods allow for high classification accuracy in noisier environments, others also exist that should be investigated in upcoming work. Particularly, wavelet decomposition can summarize an ROI by low-pass filtering and down-sampling such that only the most distinctive features are represented. This method can allow for the peaks and valleys to be represented as features, at the cost of more features in the final data set. As there is no closed-form solution to how much undersampling in this manner can be tolerated while still being useful to device authentication, more research in this area is needed.

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Air Force, U.S. Department of Defense, or the U.S. Government.

REFERENCES


