Research Article

Securing ZigBee Commercial Communications Using Constellation Based Distinct Native Attribute Fingerprinting

Christopher M. Rondeau, J. Addison Betances, and Michael A. Temple

Department of Electrical and Computer Engineering, US Air Force Institute of Technology, Wright-Patterson AFB, Dayton, OH 45433, USA

Correspondence should be addressed to Michael A. Temple; michael.temple@afit.edu

Received 19 April 2018; Revised 8 June 2018; Accepted 20 June 2018; Published 11 July 2018

Academic Editor: Bela Genge

Copyright © 2018 Christopher M. Rondeau et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This work provides development of Constellation Based DNA (CB-DNA) Fingerprinting for use in systems employing quadrature modulations and includes network protection demonstrations for ZigBee offset quadrature phase shift keying modulation. Results are based on 120 unique networks comprised of seven authorized ZigBee RZSUBSTICK devices, with three additional like-model devices serving as unauthorized rogue devices. Authorized network device fingerprints are used to train a Multiple Discriminant Analysis (MDA) classifier and Rogue Rejection Rate (RRR) estimated for 2520 attacks involving rogue devices presenting themselves as authorized devices. With MDA training thresholds set to achieve a True Verification Rate (TVR) of TVR = 95% for authorized network devices, the collective rogue device detection results for SNR ≥ 12 dB include average burst-by-burst RRR ≈ 94% across all 2520 attack scenarios with individual rogue device attack performances spanning 83.32% < RRR < 99.81%.

1. Introduction

The need to establish reliable and secure communications remains a challenge across commercial Industrial Internet of Things (IIoT) applications that support Critical Infrastructure (CI) elements (water treatment, petroleum product distribution, transportation, etc.) that are commonly operated through Industrial Control System (ICS) architectures. ZigBee networks are common within the IIoT and CI/ICS domains and remain a mainstay for implementing wireless sensor and automation networks supporting medical, smart home and building automation, and consumer electronics [1–3]. The degree of required ZigBee antihacking security varies with application criticality and will increase as the number of deployed ZigBee devices under 802.15.4 market expansion grows to 1 billion units being shipped annually by 2022 and the next generation multiprotocol 802.15.4/Bluetooth/WiFi hardware becomes available [4]. As device makers strive to take advantage of market opportunity and satisfy consumer wants for the next “greatest” interface device, it remains unclear that they have taken necessary prudent steps to address legacy security concerns.

In light of vital asset vulnerability, protection of IIoT CI and ICS elements has become a national-level priority for both the public and private sectors [5–7]. Mitigation strategies against cyberattacks have traditionally focused on bit-level solutions targeting the higher communication protocol layers and until recently there has been minimal emphasis on physical (PHY) layer development [8–10]. This work addresses hardware device identity (ID) verification as a means to enhance network security by preventing unauthorized access through the PHY doorway through which a preponderance of malicious cyberattacks occur. The focus on ZigBee device security is motivated by two factors, including the following: (1) ZigBee and related 802.15.4 communication systems are deployed world-wide and (2) ZigBee serves as a representative protocol for broader IIoT applications [11, 12]. This work expands previous wireless device ID discrimination activity that has successfully exploited various Distinct Native Attribute (DNA) features extracted from selected signal responses to reliably discriminate transmitting hardware devices.

The Constellation Based DNA (CB-DNA) development here is motivated by concepts introduced in [13] used to
2. Background

2.1. Quadrature Amplitude Modulation (QAM). The general development for the class of complex M-ary QAM modulated signals having in-phase/quadrature-phase (I/Q) components includes the mth complex data modulated symbol given by

$$S_m(t) = I_{n_m} + jQ_{n_m},$$  \hspace{1cm} (1)

for 0 < t < T_{sym}, where T_{sym} is the total symbol duration, m = 1, 2, . . . , M, and I_{n_m} and Q_{n_m} are real-valued modulation components in the I/Q constellation space with I_{n_m} \in \{I_0, I_1, \ldots , I_{N_m}\} and Q_{n_m} \in \{Q_0, Q_1, \ldots , Q_{N_m}\}. For complex symbols given by (1), a transmitted (Tx) burst of N_{sym} QAM modulated symbols is given by

$$S_{tx}(t) = \sum_{m=1}^{N_{sym}} S_m(t-kT_{sym}) \exp(2\pi f_c t + \phi_{tx}),$$

$$S_{tx}(t) = \sum_{m=1}^{N_{sym}} S_m(t-kT_{sym}) \cos(2\pi f_c t + \phi_{tx}),$$  \hspace{1cm} (2)

$$+ j \sum_{m=1}^{N_{sym}} S_m(t-kT_{sym}) \sin(2\pi f_c t + \phi_{tx}),$$

for 0 < t < N_{sym} \times T_{sym} with f_c being the transmitted carrier frequency and \phi_{tx} = \phi/2 accounting for quadrature-phase error induced by hardware components [21]. The sequence of ideal transmitted QAM symbols in S_{tx}(t) is denoted by vector S_m = (S_1, S_2, \ldots , S_M, S_{N_{sym}-1}, S_{N_{sym}}) where S_m \in \{S_1, S_2, \ldots , S_M\}. For the case of M = 4-ary signaling, the QAM S_{tx}(t) expression in (2) can be used to effectively represent the 4-ary Offset Quadrature-Phase Shift Keyed (O-QPSK) used here for ZigBee demonstration.

Considering channel amplitude A_{ch} and transmitter-to-receiver propagation delay \tau_{ch} factors, the received (Rx) burst corresponding to S_{tx}(t) in (2) is given by

$$S_{rx}(t) = A_{ch} S_{tx}(t - \tau_{ch}),$$  \hspace{1cm} (3)

which has baseband received I_{rx}(t) and Q_{rx}(t) components that can be expressed as

$$I_{rx}(t) = G_{I/Q} \sum_{k=1}^{N_{sym}} I_{S_k} \left( t - kT_{sym} - \tau_D \right) + O_{I}(t),$$  \hspace{1cm} (4)

$$Q_{rx}(t) = G_{I/Q} \sum_{k=1}^{N_{sym}} Q_{S_k} \left( t - kT_{sym} - \tau_D \right) + O_{Q}(t),$$  \hspace{1cm} (5)

where G_{I/Q} is the I/Q gain imbalance, \tau_D accounts for \tau_{ch} and relative time delay between receiver I/Q channels, and O_{I}(t) and O_{Q}(t) represent I/Q offset factors [21]. The G_{I/Q}, \tau_D, O_{I}(t), and O_{Q}(t) factors in (4) and (5) collectively account for transmit and receive imperfections. The sequence of corrupted received QAM symbols in S_{rx}(t) is denoted by vector S_k = (S_1, S_2, \ldots , S_M, S_{N_{sym}-1}, S_{N_{sym}}).

The cumulative effect of transmitter-receiver imperfections and channel errors captured in I_{rx}(t) and Q_{rx}(t) components is a degradation in received QAM symbol estimates, denoted here as S_k = (S_1, S_2, \ldots , S_M, S_{N_{sym}-1}, S_{N_{sym}}) for a given S_k induced by a location shift of received C_k = I_k + jQ_k QAM constellation points relative to the corresponding ideal transmitted C^0_k = I_k + jQ_k constellation points. In addition to potential QAM symbol estimation error induced by received C^0_k deviation, there are two other receiver processes that are key for achieving reliable QAM symbol estimation, including (1) received carrier frequency offset f_{rx} estimation and (2) phase recovery for symbol constellation derotation.

2.1.1. Received Carrier Estimation. Following downconversion by f_c and baseband filtering, samples of the received M-QAM signal at the receiver’s Matched Filter (MF) output can be modeled as

$$S_{mf}(n) = K_R S_k(n) \exp(j2\pi f_{rx} n) + N_f(n),$$  \hspace{1cm} (6)

where n = 1, 2, \ldots , N_{mf}, K_R is a real-valued scalar, S_k are the transmitted QAM symbols in (2), f_{rx} is relative received carrier frequency offset, and N_f is communication channel background noise [22]. The residual f_{rx} in S_{rx}(t) can be estimated by raising S_{mf}(n) in (6) to the Mth power to remove the modulation effects. This effectively creates a multitone spectral response with a dominant (highest power) tone.
occuring at \( M \times f_{Rx} \) [23]. This is illustrated for 4-QAM where \( S_{MF}^4(n) \) can be expanded as

\[
S_{MF}^4(n) = \left[ K R S_m(n) \right]^4 \exp(j8\pi f_{Rx} t) + 4 \left[ K R S_m(n) \right]^3 \exp(j6\pi f_{Rx} t) N_b(n) + 4 \left[ K R S_m(n) \right]^2 \exp(j2\pi f_{Rx} t) N_b^2(n) + 6 \left[ K R S_m(n) \right] \exp(j4\pi f_{Rx} t) N_b^2(n) + N_b^4(n)
\]

which includes a dominant \( 8\pi f_{Rx} = 2\pi(4f_{Rx}) \) frequency component. The estimated received carrier frequency offset is given by \( \tilde{f}_{Rx} = 4[\arg\max (F(S_{MF}^4(n))) \] where \( F(\cdot) \) denotes the discrete Fourier transform.

2.1.2. Constellation Phase Recovery. Receivers commonly use a Phase Locked Loop (PLL) to reconstruct the suppressed carrier via dynamic feedback that autoaccommodates for phase errors [24]. While generally beneficial, this within-burst autoaccommodation can potentially obscure subtle DNA feature differences that may help discriminate transmitters. Therefore, burst-by-burst discrete phase estimation and constellation derotation was implemented using an algorithm that rotates the received \( C^6_k \) constellation points for each burst from 0 to \( \pi/2 \) radians in \( N_A = 100 \) increments and selects the phase rotation angle yielding the minimum variance between the incrementally rotated pool of received \( C^6_k \) and the ideal \( C^6_k \) constellation points. The pseudocode for implementing this algorithm is presented in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Constellation phase derotation algorithm.</th>
</tr>
</thead>
</table>

Required: Received Constellation Projection \( C^6_k \)
RotationVariances \( \leftarrow 0 \)
for \( N_A = 1 \) to \( 100 \) do
\( \theta \leftarrow (N_A \cdot \pi)/(2 \times 100) \)
\( \text{Rot}(C^6_k) \leftarrow C^6_k \cdot e^{j\theta} \)
\( \text{Temp} \leftarrow |\text{Re} \{ \text{Rot}(C^6_k) \}| + j \text{Im} \{ \text{Rot}(C^6_k) \} \)
RotationVariances \( (N_A) \leftarrow \text{Variance}(\text{Temp}) \)
end for
\( N_A \leftarrow \arg\min_{i} \{ \text{RotationVariances} \} \)
return \( C^6_k \cdot e^{j(N_A-1)/(2\times100)} \)

There are four different phase ambiguity that can exist after derotating the constellation using the algorithm in Table 1. These are resolved using estimated rotation angles of known preamble (training) symbols. The rotated constellation projections can also be normalized by scaling (dividing) each \( \text{Rot}(C^6_k) \) point by the mean \( |\text{Rot}(C^6_k)\rangle \) which locates the center of all constellation clusters on the unit circle.

2.2. ZigBee Communications. The ZigBee Communication protocol includes a Medium Access Control (MAC) layer, where device IDs are verified using bit-level credentials, that interfaces with the RF communications channel through the PHY layer using RF hardware and firmware [25]. The PHY layer is implemented according to the IEEE 802.15.4 standard for low data-rate, low-power, and short range RF communications [20]. It is estimated that more than one billion 802.15.4 compliant components will be sold by the end of this decade with a majority of them supporting localised smart home networks [4]. One such component is the Atmel AT86RF230 radio transceiver that is hosted on RZUSBSTICK devices [26]. These are low-power devices that support ZigBee operation at 2.4 GHz through an integrated folded dipole antenna with a net peak gain of \( G_A = 0 \) dB. Accounting for \( G_A = 0 \) dB and maximum AT86RF230 output power of \( P_{O_{RF}} = +3.0 \) dBm [27], the effective transmit power of the RZUSBSTICK is \( P_{Tx} = +3.0 \) dBm which make it a viable alternative for not only smart home networks but other wireless sensor networks, industrial control system, and building automation [27]. Details for the specific RZUSBSTICK devices used for demonstration are provided in Table 2 which shows the unique ZigBee Communication (ZC) device IDs assigned for experimentation.

The use of PHY layer O-QPSK modulation is mandatory for ZigBee operation at 2.4 GHz, with the O-QPSK modulator preceded by a 4-to-32 (information bit-to-spread chip) Pseudorandom Noise (PN) mapping such that the information bits are transmitted at an effective rate of (2M Chips/Sec) \( \times (4/32 \) Bits/Chip) = 250K Bits/Sec [20, 25]. Accounting for I/Q channel offset processing in the modulator, the corresponding output O-QPSK communication symbol rate for a transmitted \( S_{Tx}(t) \) burst given by (2) is \( R_{Sym} = 1/T_{Sym} = (250 \) KBits/Sec)/(2 Bits/Sym) = 125K Sym/Sec.

The required 4-to-32 PN mapping for 2.4 GHz ZigBee operation is shown in Table 3 [20]. Given this mapping, there are specific transmitted O-QPSK \( S_m \) symbol sequences that occur with varying probability. For example, the bold highlighted \( \{1 \ 0 \ 0 \ 1 \ 0 \ 0 \} \) 6-bit pattern in the output chip sequences in Table 3 is among the most frequently occurring ones (appears in 13 of 16 chip sequences) and produces the O-QPSK transmitted symbol sequence \( S_m = (S_2, S_3, S_3, S_5, S_6) \).

A 5-symbol \( S_m \) vector is denoted in Table 4 by an * and
Table 3: Input-output sequences for ZigBee 4-to-32 premodulation PN mapping [20]. Bold entries highlight one of 30 highest probability 6-bit sequences.

<table>
<thead>
<tr>
<th>Input [b_n, b_{n-1}, b_2, b_1]</th>
<th>Output Chip Sequence [c_0, c_1, c_2, ..., c_3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>11010001100011010110010011010011010110110</td>
</tr>
<tr>
<td>0100</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
<tr>
<td>0110</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
<tr>
<td>1000</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
<tr>
<td>1110</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
<tr>
<td>0000</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
<tr>
<td>0100</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
<tr>
<td>0110</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
<tr>
<td>1000</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
<tr>
<td>1110</td>
<td>11010000110100110001100001010000001010101</td>
</tr>
</tbody>
</table>

Table 4: 30 highest probability 5-symbol S_m for Table 3 mapping with * denoting S_m for the output bit sequence [1 0 0 1 0 0] highlighted in Table 3.

<table>
<thead>
<tr>
<th>(S_1, S_2, S_3, S_4, S_5)</th>
<th>(S_1, S_2, S_3, S_4, S_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
<tr>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
<td>(S_1, S_2, S_3, S_4, S_5)</td>
</tr>
</tbody>
</table>

is among the 30 highest probability transmitted O-QPSK S_m used for conditional CB-DNA demonstration.

2.3. Device Classification and Device ID Verification. Device discrimination (classification and ID verification) is performed using DNA fingerprints with a Multiple Discriminant Analysis/Maximum Likelihood (MDA/ML) process adopted from [11]. This includes MDA model training for N_{C_{d}} classes (ZC devices) with components of (1) an N_TST x N_{C_{d}}-1 dimensional fingerprint vector (F) into the N_{C_{d}}-1 dimensional space containing fingerprint projection P_F = FW; (2) an N_{C_{d}} dimensional fingerprint scaling vector α; and (3) the N_{C_{d}} training means (μ) and covariances (Σ). MDA models are generated using a pool of 4400 total fingerprints per class that are equally divided into N_{TNG} = 2200 Training (even indexed fingerprints) and N_{TST} = 2200 Testing (odd indexed fingerprints) subsets. The even-odd indexing assignment ensures the models account for temporal channel variation, collection bias, etc., effects occurring during the course of emission collection.

The TNG fingerprints at a given SNR are used for MDA model training that includes K = 5-fold cross-validation [Dudlík] with the best projection matrix W_{Best} selected as the fold producing the highest cross-validation accuracy. The TST fingerprints are then input to the model and a 1 versus N_{C_{d}} best match ML classification decision is made based on a selected classification test statistic (Z_{C_{d}}). The trained classification yielding highest conditional probability P(Z_{C_{d}} | Z_{C_{d}}) for all i = 1, 2, ..., N_{C_{d}} is the called class (right or wrong) assigned to the unknown input fingerprint F. Classification performance at a given SNR is presented in an N_{C_{d}} x N_{C_{d}} confusion matrix, with (1) average cross-class percent correct classification (%C) calculated as the sum of diagonal (correct) matrix entries divided by the total number of classification trials (N_{C_{d}} x N_{TST}) and (2) individual class %C for each class C_i calculated as the sum of ith row entries divided by N_{TST}. Alternately, classification performance is presented in %C versus SNR plots.

The device ID verification process uses the selected MDA model components (W, α, μ, and Σ) and device TST fingerprints to estimate both (1) authorized network device True Verification Rate (TVR) (true positive) and (2) unauthorized device Rogue Rejection Rate (RRR) (true negative). For a given claimed (unknown) authorized device ID to be verified, the process includes the following: (1) projecting TST F fingerprints for the device under test into the N_{C_{d}}-1 discrimination space using P_F = μ ⊗ FW where ⊗ denotes element-by-element vector multiplication, (2) calculating the selected verification test statistic (Z_{TVR}) for N_{TST} total fingerprints using training μ and/or Σ for the claimed authorized device ID, (3) forming a normalized (unit area) Probability Mass Function (PMF) using N_{TST} total Z_{TVR}, (4) overlaying a desired training verification threshold (t_{TVR}), and (5) calculating the PMF area above/below t_{TVR} to achieve the desired verification rate. Common Z_{TVR} measures of similarity include (1) distance-based metrics such as the Euclidean distance between projected P_F and the claimed training class mean μ and (2) probability-based metrics that map the calculated P_F Euclidean distance to a normalized multivariate Gaussian probability distribution having mean μ and covariance Σ. Euclidean distance is perhaps the most easily conceptualised and was chosen here for proof-of-concept demonstration.

The PMFs in Figure 1 are used to illustrate Device ID verification for Euclidean distance "lower-is-better" measure of similarity [11]. Given these PMFs, the ID verification process includes (1) using network ZC TNG fingerprint Z_{TVR} to set the training verification threshold t_{TVR} shown in Figure 1(a) to achieve the desired TVR (blue PMF area) where PMF is for ZC TNG and PMF2 is based on accumulated TNG Z_{TVR} for
Figure 1: PMFs showing device dependent \( t_V(i) \) set to achieve desired network \( ZC_i \) TVR (true positive) given by blue PMF1 area in (a) and resultant RRR (true negative) for \( ZR_j \) device given by blue PMF2 area in (b) [11].

![Graph showing PMFs for device verification](image)

**Figure 2:** ROC curves for Figure 1 PMFs with indicated operating points based on desired TVR = 90%.

![Graph showing ROC curves](image)

3. CB-DNA Fingerprinting Development

Time domain RF-DNA Fingerprinting has historically exploited statistical features extracted from partial-burst responses where *invariant* (data independent) synchronisation and channel estimation (preamble, midamble, etc.) symbols are transmitted [15, 28–30]. The CB-DNA Fingerprinting method developed here differs considerably and exploits features extracted from full-burst responses, including regions where *variant* (data dependent) symbols are transmitted. The CB-DNA Fingerprinting development here is motivated by concepts first used in [13] to discriminate all "other" network \( ZC_k \) \((k = 1, 2, \ldots, N_{CZ}, \text{and } k \neq i)\) and (2) calculating the corresponding RRR (true negative, blue PMF2 area) in Figure 1(b) where PMF1 is the same and PMF2 is based on TST fingerprint \( Z_V \) for the rogue \( ZR_j \) device. ID verification performance can be based on TNG \( t_V(i) \) set for either (1) equal error rate conditions with False Verification Rate (FVR) given by \( FVR = 1 - TVR \) or (2) a specific desired authorized TVR.

The authorized TVR (true positive) versus FVR (false positive) trade-off is effectively captured in a Receiver Operator Characteristic (ROC) curve [Faw1] as shown in Figure 2 using Figure 1 PMFs with varying the TNG verification threshold \( t_V \) varied from \( \text{Min}[Z_V] \) to \( \text{Max}[Z_V] \). Figure 2(a) shows TVR versus FVR with the indicated operating point (●) corresponding to desired TVR = 90% and yielding \( FVR \approx 1.2\% \). Figure 2(b) shows TVR versus RAR where Rogue Accept Rate (false positive) is used to estimate the RRR ≈ 1-RAR shown along the x-axis for three arbitrary ZR devices (▼, ▲, and ■) and the TVR = 90% operating point.
Ethernet cards but it fundamentally differs in that work in [13] is based on features extracted from a contrived (nonconventional) binary constellation while the development here is for any application using conventional M-QAM signaling as introduced in Section 2.1. The development for unconditional and conditional fingerprinting is supported by the process depicted in Figure 3.

For ideal transmitted symbols having constellation projections $C^S_m$ such as those shown in Figure 3(a), the $k$th received QAM symbol in burst $S_{Rx}(t)$ of (3) is denoted as $S_k$ for $t_k < t < t_k + T_{Sym}$ where $t_k$ is the symbol start time, $T_{Sym}$ is the symbol duration, and $k = 1, 2, \ldots, N_{Sym}$ where $N_{Sym}$ is the total number of symbols in a received burst. Following synchronisation to the $k$th symbol interval, the QAM receiver extracts symbol $S_k$ and projects it to a single point $C^{Sk}_k$ in the QAM constellation space (Figure 3(b)). The corresponding estimated transmitted symbol is determined as $\hat{S}_k = S_m : \arg\min_m |C^{Sk}_k - C^S_m|$ for $S_m \in \{S_1, S_2, \ldots, S_M\}$ (Figure 3(c)). For generating unconditional CB-DNA statistical fingerprint features, the $N_{Sym}$ received $C^{Sk}_k$ in each $S_{Rx}(t)$ burst are grouped based on their corresponding $\hat{S}_k = S_m$ estimate with the group of $C^{Sk}_k$ yielding the $m$th QAM symbol estimate denoted by the sequence $\{C^{Sk}_m\}$ for $m = 1, 2, \ldots, M$.

While some prior works have investigated constellation error differences as a means for device discrimination [31], e.g., mean and variance, of Euclidean distances between received $C^{Sk}_k$ and ideal $C^{Sk}_k$, the approach here exploits constellation spatial statistical differences in $\{C^{Sk}_m\}$ groups which are induced by channel propagation and hardware variability (e.g., I/Q imbalance) resulting from component differences (oscillator phase noise, spurious mixer tones, manufacturing processes, etc.) [21]. The exploitation of these differences was first demonstrated for the contrived binary constellation work in [13] which showed that the statistical distribution of $\{C^{Sk}_m\}$ elements around the corresponding ideal $C^S$ point is conditional, i.e., the location of a given $C^{Sk}_k$ in the received QAM constellation space is dependent upon symbols received just prior to and immediately following $S_k$; these two symbols are denoted as $S_{k-1}$ and $S_{k+1}$, respectively.

The device discrimination improvement in [13] using conditional fingerprint features from the contrived binary constellation motivated formal development of the multisymbol constellation conditioning (subgrouping) method for M-QAM signaling. For the $S_k$ dependent $\{C^{Sk}_m\}$ group sequences, the basic process includes considering multiple consecutive received QAM symbols in a $S_{Rx}(t)$ burst which are denoted here by vector $S_k = (\ldots, S_{k-2}, S_{k-1}, S_k, S_{k+1}, S_{k+2}, \ldots)$ where $S_k$ is the central reference symbol. These received symbols have corresponding estimates that are used to form vector $\hat{S}_k = (\ldots, \hat{S}_{k-2}, \hat{S}_{k-1}, \hat{S}_k, \hat{S}_{k+1}, \hat{S}_{k+2}, \ldots)$ where $\hat{S}_k$ is the estimate for reference symbol $S_k$. Multisymbol constellation conditioning involves parsing each of the unconditional $\{C^{Sk}_m\}$ groups into conditional $\{C^{Sk}_m\}$ subgroups for $n = 1, 2, \ldots, N_{Sk}$ total subgroups with $S_{kn}$ denoting the $n$th subgroup. The parsing of unconditional $\{C^{Sk}_m\}$ sequences and selection of $N_{Sk}$ subgroups is somewhat arbitrary but performed with a goal of maximising cross-subgroup distribution differences that will be captured in statistical fingerprint features.

![Figure 3: Illustration of unconditional and conditional 4-QAM constellation processing.](image-url)
The subgrouping of \( \{ C_m^{\text{cb}} \} \) is illustrated (as shown in Figure 3(d)) by considering three received symbols of \( S_k = (S_{k-1}, S_k, S_{k+1}) \) and a set of \( N_{SG} \) desired subgroup conditioning vectors \( G^m \) of equivalent dimension and denoted by \( G^m = (G_1^m, G_2^m, G_3^m) \) where \( G_i^m \in [S_1, S_2, ..., S_M] \). The process for assigning each element of the \( m \)th \( \{ C_m^{\text{cb}} \} \) group to one of \( N_{SG} \) subgroups based on \( G^m \) conditions includes (1) taking each received \( S_k \) producing \( C_m^{\text{cb}} \), (3) estimating received \( \hat{S}_k \) and \( \hat{S}_{k+1} \) and forming \( \hat{S}_k = (S_{k-1}, \hat{S}_k, S_{k+1}) \), and (4) comparing the resultant \( \hat{S}_k \) with each desired \( G^m \). If \( |\hat{S}_k - G^m| \neq 0 \) for all \( n = 1, 2, ..., N_{SG} \) the \( C_m^{\text{cb}} \) under evaluation is assigned to the \( n \)th conditional \( \{ C_m^{\text{cb}} \} \) subgroup. If \( |\hat{S}_k - G^m| = 0 \) for all \( n \) the \( C_m^{\text{cb}} \) under evaluation is assigned to an "other" conditional subgroup. Formation of the two extreme phase change sequences is illustrated using Figure 4 which shows the subgrouping of \( C_m^{\text{cb}} \) sequences and conditional \( \{ C_m^{\text{cb}} \} \) sequences are used to form CB-DNA fingerprints. The construction processes for unconditional \( F_{\text{CB}UN} \) and conditional \( F_{\text{CB}CN} \) CB-DNA fingerprint vectors are identical and presented for an arbitrary complex sequence \( X \) having \( N_X \) elements. The fingerprint statistics are calculated using (1) polar magnitude \( \{ \text{Mag}(X) \} \) and angle \( \{ \text{Ang}(X) \} \) components and (2) rectangular real \( \{ \text{Re}(X) \} \) and imaginary \( \{ \text{Im}(X) \} \) components of \( X \). While any number of statistics could be used, the specific CB-DNA features used for polar representation include variance \( \{ \sigma^2 \} \), skewness \( \{ \gamma \} \), and kurtosis \( \{ \kappa \} \) statistics of both the magnitude \( \{ \text{Mag}(X) \} \) and angle \( \{ \text{Ang}(X) \} \) sequences for a total of \( 6 \) polar statistics. For the rectangular \( \{ \text{Re}(X), \text{Im}(X) \} X \) matrix representation, the calculated statistics include three unique covariance\( \sigma^2 \{ \gamma \} \) values, two nontrivial coskewness moments \( \gamma \{ \} \), and three nontrivial cokurtosis \( \kappa \{ \} \) moments [32]. Accounting for all possible statistics, the Statistical Fingerprint vector for complex sequence \( X \) is formed as

\[
F^X = [\sigma^2_{\text{Mag}(X)} \gamma_{\text{Mag}(X)} \kappa_{\text{Mag}(X)} \sigma^2_{\text{Ang}(X)} \gamma_{\text{Ang}(X)} \kappa_{\text{Ang}(X)} \sigma^2_{\text{Re}(X)} \gamma_{\text{Re}(X)} \kappa_{\text{Re}(X)} \gamma_{\text{Im}(X)} \kappa_{\text{Im}(X)}]_{1 \times N_{stat}},
\]

where \( N_{stat} = 14 \) if all indicated statistics are included.

For unconditional CB-DNA Fingerprinting \( F^X \) in (8) is calculated for all \( m = 1, 2, ..., M \) constellation symbols with \( X = \{ C_m^{\text{cb}} \} \) and the resultant \( F_m^X \) concatenated to form the final composite unconditional CB-DNA Fingerprint vector \( F_{\text{CB}UN} \) given by

\[
F_{\text{CB}UN} = \left[ F_1^X : F_2^X : ... : F_M^X \right]_{1 \times N_N^{UN}},
\]

where \( N_N^{UN} = N_{stat} \times M \) is the total number of unconditional CB-DNA features.

For conditional CB-DNA Fingerprinting \( F^X \) in (8) is calculated for all \( n = 1, 2, ..., N_{SG} \) subgroups of each \( m = 1, 2, ..., M \) constellation symbol using \( X = \{ C_m^{\text{cb}} \} \). The resultant \( F_{\text{CB}CN} \) vectors are used form the \( m \)th Conditional CB-DNA Fingerprint vector \( F_m^{\text{CN}} \) given by

\[
F_m^{\text{CN}} = \left[ F_{SG(n,1)}^{\text{CN}} : F_{SG(n,2)}^{\text{CN}} : ... : F_{SG(n,N_SG)}^{\text{CN}} \right]_{1 \times (N_{stat} \times N_{SG})},
\]
Figure 4: Received unconditioned QAM constellation at SNR = 12 dB for a burst of $N_{sym} \approx 3400$ symbols producing approximately 850 total projections in the indicated $C^S_k$, $C^S_2$, $C^S_3$, and $C^S_4$ quadrant groups.

Figure 5: Received conditional QAM constellation points for $S_1$ quadrant projections in Figure 4 showing pairwise relationship of five conditioned $\{C^S_k\}$ subgroups with elements assigned using the indicated $G^n$ conditions.
which are concatenated for all \( m = 1, 2, \ldots, M \) to form the composite conditional CB-DNA Fingerprint vector

\[
\mathbf{f}_C^{\text{CND}} = [\mathbf{f}_1^{\text{CND}} ; \mathbf{f}_2^{\text{CND}} ; \cdots ; \mathbf{f}_M^{\text{CND}}]_{1 \times N_F^{\text{CND}}},
\]

where \( N_F^{\text{CND}} = N_{\text{Stat}} \times N_{\text{SG}} \times M \) is the total number of conditional CB-DNA features. In general, unconditional and conditional CB-DNA fingerprint features can be generated using all or a subset of noted statistics, calculated for all or a subset of available projected \( \{ \mathbf{C}_m^{\text{Stat}} \} \) groups or \( \{ \mathbf{C}_m^{\text{Stat}} \} \) subgroups. The choice of which statistics and which groups to use may vary with the specific communication application (fixed, mobile, urban, city, etc.) and determines the final number of \( N_U^{\text{CB}} \) and \( N_C^{\text{CB}} \) features generated.

4. CB-DNA Fingerprinting Demonstration

ZigBee transmissions were collected for all RZUSBSTICK devices listed in Table 2 using an X310 Software Defined Radio (SDR) having an RF bandwidth of \( W_{RF} = 10 \) MHz and operating at a sampling rate of \( f_s = 10 \) MSPs in both the I/Q channels. Subsequent postcollection signal processing was performed using MATLAB and included burst-by-burst (1) center frequency estimation, (2) baseband (BB) downconversion and filtering using a 16th-order Butterworth filter having a -3 dB bandwidth of \( W_{BB} = 2 \) MHz, (3) constellation phase derotation, and (4) unconditional and conditional CB-DNA fingerprint generation per Section 3. The CB-DNA fingerprints were used to generate demonstration results for a total of \( N_{NC} = 10 \) -choose-3 = 120 unique network configurations with the \( N_{ZR} = 3 \) chosen devices serving as unauthorized attacking ZigBee Rogue (ZR) devices and the remaining \( N_{ZC} = 7 \) devices serving as authorized ZC network devices.

For each network configuration, the RRR was estimated for the \( N_{ZR} = 3 \) rogue devices using the device ID verification process detailed in Section 2.3. For each network configuration, each of the \( N_{ZR} = 3 \) ZR devices presents false ID credentials for all \( N_{\text{CZ}} = 7 \) authorized ZC network devices for a total of \( 7 \times 3 = 21 \) ZRj:ZCi assessments per network configuration. Considering all networks, a total of \( 120 \times 21 = 2520 \) ZRj:ZCi device ID verification (rogue detection) assessments were completed. Alternately, each ZC device in Table 2 served as an attacking ZR device 36 times for a total of \( 36 \times 7 = 252 \) ZRj:ZCi device ID verification assessments per RZUSBSTICK device. The RRR estimates are based on a total of 4400 fingerprints per ZR device that are presented on a fingerprint-by-fingerprint basis for ID verification; the assessments here do not include nor account for envisioned benefits to be realised by averaging fingerprints, features, etc., prior to making a final authorized versus rogue verification decision. For presentation brevity, limited results are presented herein that are representative of the poorest (lowest RRR) and best (highest RRR) results obtained across all \( N_{NC} = 120 \) network configurations and are sufficient for supporting proof-of-concept demonstration conclusions.

4.1. Authorized Network Device Classification. Device classification is first required to generate the MDA/ML models (\( \mathbf{W}, \alpha, \mu, \) and \( \Sigma \)) required for device ID verification. The CB-DNA Fingerprinting results in Figure 6 were generated using unconditional and conditional features for all \( N_{NC} = 120 \) networks. Results show \( %C \) versus SNR for all 120 networks along with cross-network average \( %C \) (solid lines) and extreme bounds (dashed lines with \( \circ \) markers) for highest and lowest \( %C \). The benefit of constellation conditioning is evident by comparing cross-network averages which show that the \( %C \approx 90\% \) benchmark is achieved for conditional features (\( \bullet \)) at SNR = 11 dB and unconditional features (\( \triangle \)) at SNR = 14 dB. For presentation brevity, additional results in this section are presented for conditional CB-DNA Fingerprinting only given its superiority.

For conditional CB-DNA Fingerprinting at SNR = 12 dB in Figure 6(b), the extreme results include (1) lowest \( %C \approx 86.78\% \) performance for Model \#1 (excludes ZC1, ZC2, and ZC3 devices) and (2) highest \( %C \approx 98.75\% \) performance for Model \#90 (excludes ZC4, ZC5, and ZC10 devices). The classification confusion matrices for these extreme cases are provided in Tables 5 and 6 and suggest that the inclusion of ZC4, ZC5, ZC6, and ZC10 devices in Model \#1 is most detrimental (italic entries in Table 5). Of note from Table 2 is that package markings for the ZC2, ZC3 pair differs from all other package markings. Thus, Model \#1 versus Model \#90 performance is consistent with historical DNA discrimination given that the ZC2, ZC3 pair is (1) excluded in the poorest Table 5 results (model includes all like-model, similarly marked devices) and (2) included in the highest Table 6 results (model includes a higher number of like-model dissimilarly marked devices).

4.2. Authorized Network Device ID Verification. SNR dependent MDA/ML model components (\( \mathbf{W}, \alpha, \mu, \) and \( \Sigma \)) from Section 4.1 are used to assess authorized network ZC device ID verification at selected verification SNR\(_V\). Results are presented for conditional CB-DNA fingerprints at SNR\(_V\) = 12 dB where average MDA/ML performance in Figure 6(b) achieves the \( %C \approx 90\% \) benchmark. For each network, device TNG fingerprints are used to set device dependent \( t_V(i) \) for all authorized devices to achieve TVR \( \approx 95\% \). \( t_V(i) \) for the worst and best performing MDA/ML models in Figure 6(b) are shown in Figure 7(a) (Model \#1) and Figure 7(b) (Model \#90). \( t_V(i) \) are overlaid with Euclidean distance TNG statistics (\( Z_V \)) and ID verification identified as either accept (\( \circ \)) or reject (\( \times \)) decisions. The accept/reject decisions and final performance are based on \( Z_V \) for N\(_{TV}\) = 2200 fingerprints per authorized device with \( Z_V < t_V(i) \) (\( \circ \) markers) representing correct ID verification (proper access granted) and \( Z_V > t_V(i) \) (\( \times \) markers) representing incorrect ID verification (improper access denial). The resultant TVR for individual ZC devices is shown along the x-axis and yields an overall cross-ZC average TVR = 94.84\% for both models.

4.3. Unauthorized Rogue Device Detection. Accounting for all \( N_{NC} = 120 \) network configurations with each of the \( N_{ZR} = 3 \) held-out ZR (\( j = 1, 2, \ldots, 10 \), \( j \neq i \)) devices serving in an attacking ZRj:ZCi role a total of 252 times (including multiple
attacks against a given ZCi device present in multiple networks), the cumulative per ZRj RRR performance averaged across all networks for $8 \leq \text{SNR}_V \leq 20$ dB is shown in Table 7. Of note here is the average cross-ZRj RRR $\approx 89.42\%$ at $\text{SNR}_V \approx 12$ dB which is approximately the same SNR where MDA/ML device classification in Figure 6(b) achieves the $\%C = 90\%$ benchmark. As shown in Table 7 $\text{SNR}_V = 12$ dB results, the lowest RRR occurs for ZR4 and ZR6 devices and the highest RRR occurs for ZR1 and ZR3 devices. Excluding $\text{SNR}_V = 8$ dB performance, collective rogue device results for $\text{SNR}_V \geq 12$ dB include (1) cumulative cross-ZR RRR $\approx 94\%$ across all ZR:ZC attack scenarios and (2) individual cross-ZR performance across 252 attacks spanning $83.32\% < \text{RRR} < 99.81\%$.

Table 5: Confusion matrix for lowest performing Model #1 in Figure 6(b) at $\text{SNR} = 12$ dB with $\%C \approx 86.78\%$ (sum of diagonals divided by 15,400 trials) and italic to highlight the largest error contributors (ZC4, ZC5, ZC6, and ZC10).

<table>
<thead>
<tr>
<th>CALLED CLASS</th>
<th>ZC4</th>
<th>ZC5</th>
<th>ZC6</th>
<th>ZC7</th>
<th>ZC8</th>
<th>ZC9</th>
<th>ZC10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZC4</td>
<td>1868</td>
<td>152</td>
<td>64</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>ZC5</td>
<td>152</td>
<td>1700</td>
<td>236</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>108</td>
</tr>
<tr>
<td>ZC6</td>
<td>128</td>
<td>232</td>
<td>1608</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>220</td>
</tr>
</tbody>
</table>

Table 6: Confusion matrix for highest performing Model #90 in Figure 6(b) at $\text{SNR} = 12$ dB with $\%C \approx 98.75\%$ (sum of diagonals divided by 15,400 trials) and italic to highlight the largest error contributors (ZC6 and ZC7).

<table>
<thead>
<tr>
<th>CALLED CLASS</th>
<th>ZC1</th>
<th>ZC2</th>
<th>ZC3</th>
<th>ZC6</th>
<th>ZC7</th>
<th>ZC8</th>
<th>ZC9</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZC1</td>
<td>2184</td>
<td>4</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZC2</td>
<td>0</td>
<td>2132</td>
<td>0</td>
<td>16</td>
<td>40</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>ZC3</td>
<td>0</td>
<td>0</td>
<td>2200</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZC6</td>
<td>12</td>
<td>24</td>
<td>0</td>
<td>2164</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZC7</td>
<td>0</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>2156</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>ZC8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2192</td>
<td>4</td>
</tr>
<tr>
<td>ZC9</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>2180</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6: Classification for 120 networks with (a) unconditional and (b) conditional CB-DNA features. Mean results show that the $\%C = 90\%$ benchmark is achieved at $\text{SNR} \approx 14$ dB (unconditional) and $\text{SNR} \approx 11$ dB (conditional).

(a) Results for $N_F^{\text{UNI}} = 36$ unconditional features with conditional mean from Figure 6(b) overlaid for comparison

(b) Results for $N_F^{\text{COND}} = 270$ conditional features with unconditional mean from Figure 6(a) overlaid for comparison
For the overall poorest ZR4 and ZR6 results in Table 7 at SNR\textsubscript{v} = 12 dB there are eight network models (#17, #45, #66, #86, #91, #92, #93, and #94) that include both ZR4 and ZR6 serving as rogue devices. Considering only these models, the cumulative ZR4 and ZR6 results include RRR ≈ 85.25% and RRR ≈ 82.03%, respectively. The overall poorest ZR4 and ZR6 RRR results for these eight models at SNR\textsubscript{v} = 12 dB are presented in Figure 8 and occur for Model #45 with ZC1, ZC3, ZC5, ZC7, ZC8, and ZC10 authorized devices. As estimated by averaging individual ZRj:ZCi RRR presented along Figure 8 x-axes, the average performance for ZR4:ZC3 is RRR ≈ 64.84% and for ZR6:ZC3 is RRR ≈ 77.56%. These are higher than the cumulative ZR4 and ZR6 results across all 120 models are presented in Table 7 and thus do not represent the overall poorest ZR4 and ZR6 device results.

For completeness, the overall poorest ZR4 and ZR6 RRR results across all 120 models are presented in Figure 9 which shows that the lowest RRR results are obtained for separate models and include average RRR = 73.27% in Figure 9(a) for ZR4 with Model #19 and average RRR = 64.84% in Figure 9(b) for ZR6 with Model #4. While it is not immediately obvious why these are the two poorest cases, these ID verification results are consistent with the increased MDA/ML classification challenge noted in Section 4.1 for models based on similarly marked authorized devices. Specifically, the poorest RRR < 80% results in Figure 9 are all attributable to ZCj:ZCi combinations of similarly marked ZC4, ZC5, ZC6, and ZC10 devices.

For the overall best RRR ZR1 and ZR3 results in Table 7 at SNR\textsubscript{v} = 12 dB there are eight network models (#1, #9, #10, #11, #12, #13, #14, and #15) that include both ZR1 and ZR3 serving as rogue devices. The overall best rogue ZR1 and ZR3 detection results for these models at SNR\textsubscript{v} = 12 dB are presented in Figure 10 and include assessments for Model #11 with ZC2, ZC4, ZC5, ZC7, ZC8, ZC9, and ZC10 authorized devices. As estimated by averaging the individual ZRj:ZCi RRR indicated along Figures 10(a) and 10(b) x-axes, the average RRR performance across best case ZR1:ZCi is RRR = 99.31% and across all ZR3:ZCi is RRR = 99.98%; this best case cross-ZRj RRR was observed for a majority of models and ZRj:ZCi considered.

5. Conclusion

An analytic development of CB-DNA Fingerprinting for conventional QAM features is presented as well as its application to verification-based rogue detection demonstrated using ZigBee RZSUBSTICK communication devices. Results are...
based on experimentally collected signals with postcollection fingerprint generation and authorized versus rogue device ID verification performed for 120 unique networks consisting of seven authorized and three unauthorized attacking rogue devices. Collective authorized device discrimination results for all 120 network configurations using an MDA classifier included (1) average cross-class percent correct classification of %C > 90% achieved for SNR ≥ 12 dB and (2) identification of device dependent verification thresholds yielding True Verification Rates (true positive) of TVR = 95% for all authorized network devices. The MDA network models were used for rogue device ID verification and Rogue Rejection Rate (RRR) (true negative) estimated for all rogues presented to the networks. Collective rogue device detection results for SNR ≥ 12 dB included (1) cumulative average burst-by-burst RRR ≈ 94% across 2520 total rogue attack scenarios and (2)
performance across 252 attacks per individual devices spanning $83.32\% < \text{RRR} < 99.81\%$. As a first successful proof-of-concept demonstration using CB-DNA Fingerprinting with conventional communication constellation features, these results are promising and further research is warranted.

Data Availability

The data used to support the findings is generally unavailable due to public releasability constraints. However, please contact the corresponding author for special release consideration.

Disclosure

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Air Force Institute of Technology, the Department of the Air Force, the Department of Defense, or the US Government. This paper is approved for public release, Case#: 88ABW-2018-2040.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References


