Inference Attacks on Physical Layer Channel State Information

Paul Walther* and Thorsten Strufe†
*Chair for Privacy and Security, TU Dresden
†Karlsruhe Institute of Technology (KIT) and Centre for Tactile Internet, TU Dresden (CeTI)
{<firstname>.<lastname>}@tu-dresden.de

***Abstract*—In Physical Layer Security, knowing the reciprocal state information of the legitimate terminals' wireless channel is considered a shared secret. Although questioned in recent works, the basic assumption is that an eavesdropper, residing more than half of a wavelength away from the legitimate terminals, is unable to even obtain estimates that are correlated to the state information of the legitimate channel.**

In this work, we present a Machine Learning based attack that does not require knowledge about the environment or terminal positions, but is solely based on the eavesdropper's measurements. It still successfully infers the legitimate channel state information as represented in impulse responses. We show the effectiveness of our attack by evaluating it on two sets of real world ultra wideband channel impulse responses, for which our attack predictions can achieve higher correlations than even the measurements at the legitimate channel.

I. INTRODUCTION

Physical Layer Security (PhySec) provides lightweight alternatives to classical cryptographic algorithms, with information theoretical proofs of security. It provides different primitives, e.g. for key agreement [28], [21], [22], or authentication [13], [25], [19]. At their core, they all rely on the assumption that the wireless channel between the legitimate parties Alice and Bob provides a common source of randomness that is secret to adversaries.

This advantage between Alice and Bob can take different forms, most commonly knowledge of the legitimate channel's characteristics. *Channel Reciprocity* dictates that the characteristics of a wireless channel between two nodes are reciprocal for the participating two terminals, if measured within some coherence time. The advantage then is based on the underlying assumption that an attacker residing more than half of a wavelength away from any of the legitimate terminals can only obtain channel estimations that are completely uncorrelated to those of the legitimate channel. The assumption itself is based on the so-called *channel diversity* or *spatial decorrelation* founded in Jakes' statistical multipath channel model [6].

Channel State Information (CSI) can be obtained as Channel Impulse Responses (CIR, cmp. Fig. 1 and 4 for examples). The legitimate communication partners mutually send impulses on the channel, which allows the other party to estimate the respective CIRs. Following the assumption above, these CIRs are highly correlated. Minor differences between those CIRs remain, due to interferences or non-reciprocal hardware paths (TX/RX), but subsequently are removed in a process called *Information Reconciliation*. The *spatial decorrelation* assumption dictates that an eavesdropping adversary observes entirely uncorrelated CIRs, which cannot be reconciled into sequences similar to the secret between Alice and Bob.

The basic premise of *spatial decorrelation* has been questioned in various works, e.g. [18], [28]. As the multipath propagation is mainly determined by the physical environment and the terminal position, the CIR might be precalculated in so called *predictable channel attacks*. In such attacks, a deterministic channel model is used to precompute the CIR between two legitimate nodes, using knowledge about the physical environment and terminal positions. Upon some theoretical work, practical attempts on real-world data have proved successful [8], [22]. They demonstrated feasibility of the attack, although with limited utility.

In this work we propose a new class of attacks that does not model the physical world explicitly, but is based on machine learning. We train a convolutional neural network (CNN) with prerecorded data. Observing the impulses of Alice and Bob at the adversary's position, this attack allows us to infer the supposedly secret CIR of their legitimate channel. This attack is fundamentally much stronger than its predecessors for the following two reasons: first, much less a priori knowledge is needed — neither concrete knowledge of the physical environment nor of the terminal positions are necessary. And second, it¸ does not require position-specific optimization, making it more generally applicable.

To evaluate this attack, we apply it in the context of *Channel Reciprocity based Key Generation* (CRKG), based on ultra-wideband (UWB) CIRs, where the reciprocal CIRs are treated as common randomness for key generation. We use UWB CIRs as their sensitivity to multipath propagation makes them very difficult for attackers to predict or infer, compared to other channel characteristics or transmission technologies. Using two data sets of real world measurements, we demonstrate that our attack can predict CIRs with cross correlations as high as those at the legitimate parties. In the context of CRKG, this yields up to 83.5% compromised key material.

This work has been supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

978-1-6654-0392-4/20/$31.00 ©2020 IEEE
DOI 10.1109/TrustCom50675.2020.00125

Our *contributions* are as follows:

- we propose a machine learning assisted attack on the secrecy of CSI in the form of CIRs
- we thoroughly evaluate this attack using two data sets of real world measurements
- we compare our results against currently best performing inference attacks
- we show the respective security implications.

The remaining paper is organized as follows: Sec. II describes necessary preliminaries for this work. Next, Sec. III gives an overview of current attacks on CRKG. In Sec. IV the adversary model is described and Sec. V presents the implementation of the attack. Finally, Sec. VI shows the performance of this attack and Sec. VII concludes the paper.

II. Background and System Model

In the following we will describe the core components of our system model: the channel model in use, the basic channel estimation protocol and since it is the motivating use case, the processing during CRKG.

For this work, we assume the usage of ultra-wideband (UWB) channel impulse respones (CIR) as channel characteristics. Hence, we follow the assumptions of the widely applied UWB multipath propagation model as defined in [6]. This means the CIR at time t, $h(t)$, is expressed as the sum of multipath components, which are caused by reflections, diffractions and scattering:

$$h(t) = \sum_{n=0}^{N} \alpha_n e^{-j\phi_n} \delta(t - t_n) \qquad (1)$$

Here, α_n and ϕ_n are the amplitude and phase of the respective n-th multipath component, and δ is the Dirac function. The properties α and ϕ are the channel characteristics used in the PhySec primitives. It is worth noting, that (1) assumes a time-invariant channel. This assumption is justified since the CIR measurements are taken within the *coherence time* of the channel, in which the channel is considered to be invariant.

The participating terminals can estimate their respective CIR by measuring the received signal $y(t)$ of a known transmitted input signal $x(t)$, since $y(t)$ is the convolution of the input signal with the channel plus Additive Gaussian White Noise n:

$$y(t) = h(t) * x(t) + n(t) \qquad (2)$$

Given the time-discrete nature of sampling, the single values within a CIR follow a defined structure as described in [17]. Accordingly, both the amplitudes of the multipath components themselves and the values within a component follow an exponential decrease. An abstract representation thereof is depicted in Fig. 1. Due to additional noise and interferences, real world measurements do not exhibit quite such a clear structure (cmp. Fig. 4a and 4c).

Using this channel model, we can describe the core **system model** for the presented attack. This system model is based on general CRKG processing, as described in [20]. It is worth noting, that although CRKG is the motivating use case here, the presented attack is not limited to this particular application — in fact, every PhySec primitive that relies on CIRs is affected, as for instance all corresponding authentication approaches, among others. The system model and core protocol for CRKG are depicted in Fig. 2.

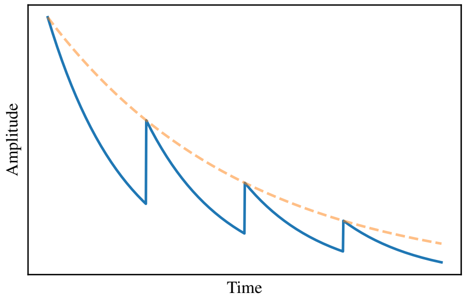


Fig. 1: Abstract structure of a CIR. The blue line represents the CIR, the orange the decay of the multipath components.

The basic workflow of CRKG is as follows: Alice and Bob exchange messages over the reciprocal channel to obtain the respective estimations of the channel characteristics h_{AB} and h_{BA}. Channel reciprocity causes their observations h_{AB} and h_{BA} to be highly correlated. Noise and interferences cause some distortion, and they hence are not perfectly equal. Therefore, subsequent to quantizing these estimates, Alice and Bob perform *Information Reconciliation*, which eliminates remaining differences between the bit vectors at Alice and Bob. They finally perform *Privacy Amplification*, which takes potential leakage to Eve into account, extracts the remaining secret randomness and yields the final key candidates.

The most important assumption of CIR-based PhySec is the claim that an adversary Eve, residing more than half of a wavelength away from the legitimate terminals (around $3cm$, in our UWB experiments), cannot estimate characteristics that are sufficiently correlated to those of the legitimate channel [15]. It

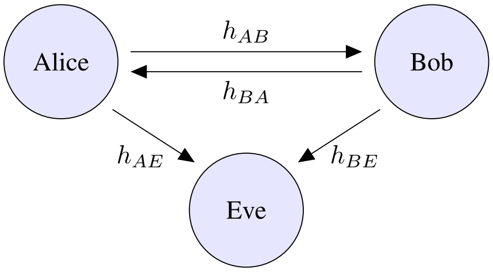


Fig. 2: Generic system model for CRKG: Alice and Bob measure the reciprocal channel and thereby obtain their estimates h_{AB} and h_{BA}. Eve overhears this communication and estimates her own channels h_{AE} and h_{BE}. Time-dependency $h_{XY}(t)$ of the estimates is included in our system model, but the notion is omitted for brevity.

is derived from *spatial decorrelation*, as conjectured by Jakes scattering theorem [6].

Since we will present an attack that tries to infer CIRs, as described in Equation (1), we need a **metric** for similarity of CIRs. Commonly, CIRs are compared using the cross correlation of the two signals [16]. This is also the common metric in related work [8], [22], and we hence will employ the *normalized cross correlation* (CC), as defined in [16]:

$$CC(h, g) = \max_k \frac{r_{gh}(k)}{\sqrt{E_g E_h}} \tag{3}$$

$$= \max_k \frac{\sum_{i=-\infty}^{\infty} g[i]h[i-k]}{\sqrt{\sum_{i=0}^{n_g-1} g[i]^2 \sum_{i=0}^{n_h-1} h[i]^2}} \tag{4}$$

The respective CIRs in comparison are denoted by h and g: Given the metric, the attacker hence tries to infer a CIR g that is as close to h_{AB} as possible (i.e. $CC(g, h_{AB})$ is close to $CC(h_{AB}, h_{BA})$ or respectively $CC(g, h_{AB}) - CC(h_{AB}, h_{BA}) \approx 0$). In other words, the attacker aims to minimize the remaining secret randomness between Alice and Bob, or preferably eliminate it completely, to break the secrecy of the key agreement.

The *received signal strength indicator* (RSSI) is another channel measurement, the use of which, besides CIR, has frequently been suggested. This is mainly due to its general availability, also in commercial off-the-shelf (COTS) radio interfaces. RSSI is a single value that indicates the received power level. Note, that knowledge of the CIR, as our adversary aims to infer, allows to directly derive the RSSI value as well [26]:

$$RSSI = 10 \log_2(||h||^2) \tag{5}$$

Considering this calculation it becomes evident that RSSI contains much less information about the channel than CIR.

Finally, we would like to point out that different PhySec primitives have different underlying requirements for the respective input data: CRKG, for example, requires dynamic, changing channel characteristics, since the entropy of these dynamic changes is used for key generation. PhySec authentication, on the other hand, requires comparatively static channel characteristics to verify the identify of the legitimate communication partner, which makes inference even easier.

III. Related Work

In the following we will describe current attacks on CRKG systems. As with the general CKRG approach, the majority of related work here is targeted at RSSI based systems.

Several active attacks against RSSI based CRKG schemes have been proposed [4], [27], [10]. At their core they all either employ jamming or signal injection, intending to deteriorate or alter the transmission in such a way, that only predictable channel characteristics are used by Alice and Bob. By forcing the usage of predictable keying material, the respective key derivation is compromised. Due to their active nature, such attacks can easily be detected and mitigated.

In the realm of passive attacks, many papers assume that the basic premise of spatial decorrelation does not hold in practice [18], [28], [29], [30]. For RSSI values it was shown, that distances greater than $\lambda/2$ do not yield completely uncorrelated channel measurements [29]. Further, a passive eavesdropper can derive up to 74.11% key bits correctly, when RSSI based schemes are used [5].

Recent work claims that CIR based PhySec is immune to *predictable channel attacks* and to the attacks presented above [14]. However, in spite of such claims, there have been practical attempts using real world measurements, which directly attack CIRs as channel characteristics in a passive setting.

Döttling et al. propose the so called *room reconstruction attack* [2] . The core idea is the following: within Eves estimated CIR are the respective multipath components, which mainly originate from reflections. The position of the these components within the CIR is the respective time delay of a multipath component. Based on this time delay and the propagation speed of the wireless wave, i.e. the speed of light, the causing reflector has to be in fixed distance. This fixed distance defines an ellipsis on which the causing reflector has to lie. As Eve overhears two CIRs, one from Alice, one from Bob, two such ellipses can be constructed. The intersection has to be the original reflector in the physical environment. The legitimate CIR may subsequently be calculated from the environment reconstructed in this way. The authors realized their idea in theory and with simulation data, but without any evaluation.

Hamida et al. and Walther et al. extend this approach, bridging the gap to practical realization [8], [22]. They aim to reconstruct the legitimate CIR between Alice and Bob, given perfect knowledge of their positions and surrounding environment. While the former use an unspecified ray-tracing tool, the latter calculate the CIRs using the deterministic UWB channel model of Kunisch and Pamp [11]. The parameters of the channel model were optimized for the given positions and environments, both specifically for each individual position as well as generic for all possible positions. While the generic attack cannot be considered a success, attacks based on individual optimization yielded somewhat promising results.

IV. Attack Idea and Adversary Model

Next, we will introduce the basic idea behind our attack and the resulting attacker model.

The **attack idea** is rooted in the following three observations regarding CIR based PhySec and CRKG: *First*, the CIRs recorded by the respective participants have an inherent structure, in the sense that the individual features follow a well-defined pattern [6], [17], as visualized in Fig 1. These are exactly the features used as key material in CRKG, for example. As [22] has demonstrated, CIR features can be estimated with deterministical calculation to a certain extent, given sufficient information. *Second*, the core assumption, that an eavesdropper who resides more than $\lambda/2$ away from the legitimate partners measures CIRs that lead to completely uncorrelated channels estimates, might not hold in practice. This assumption has already been questioned in recent work [18], [28], and in theory it is possible to reconstruct parts of the room geometry solely from overheard features [2]. Hence, we postulate that it may be possible to reconstruct parts of the key material only from features overheard in larger distances. *And finally*, as the CRKG processing itself corrects certain differences between input data during *Information Reconciliation*, an attack does not need to predict the input data perfectly. For a successful attack it is sufficient to predict input values which do not have greater differences than the respective h_{AB}, h_{BA} pair. In this work, we hence want to investigate to which extent these three facts combined facilitate a practical inference attack.

Our attack aims to predict the CIR accurate enough to be within a distance that subsequently is corrected successfully by *Information Reconciliation*. To achieve this, we propose a machine learning model, which implicitly performs geometry reconstruction as well as the subsequent prediction of legitimate channel characteristics. It allows an attacker to predict the input values for the CRKG processing, which in turn enables her to derive the supposedly secret key of Alice and Bob.

The **adversary model** is defined by the attacker's behaviour within the system model. To realize the described attack idea, the adversary acts in two steps: first, in a preparatory step, the attacker collects CIR samples representing all three channels, i.e. possible channels between Alice/Bob, Alice/Eve, and Bob/Eve. These samples are used to train our machine learning model. Second, during the attack itself, the adversary is completely passive and is only listening to the legitimate communication between Alice and Bob. The attacker locally processes the CIRs she overhears with the trained model and infers the CIR between Alice and Bob. Since all messages of the subsequent processing are sent in clear text, the attacker can subsequently use the inferred CIR to derive the same key as Alice and Bob. As the attack is carried out with COTS hardware, no special capabilities are needed for the adversary.

This attacker model is similar to those in [22], [2] and [8]. Nevertheless, we weaken the attacker model, and thereby strengthen the attack, as we do not require Eve to know the room geometry and positions of the legitimate terminals within the room. Following the notation presented in [24] this changes the adversary model from *model-based* to *model-free*, as we no longer need a predefined physical model of room and positions, but instead learn them implicitly, based on prior observations of the adversary. Further, in difference to the state of the art, our attack does not require any optimization with respect to the terminal positions. Finally, to the best of our knowledge and with respect to PhySec, it represents the first attack leveraging machine learning.

V. Realization

In the following we describe how we implemented the presented attack. First, the acquired data and the respective preprocessing is presented. Second, we will describe the core architecture of our machine learning approach. And finally, we show how these parts are combined to successfully mount the corresponding attack.

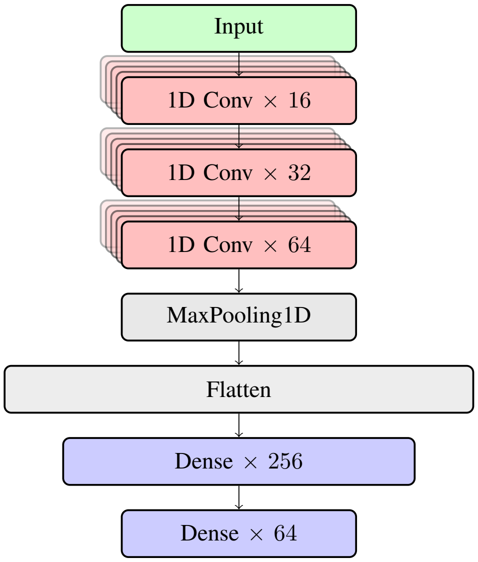


Fig. 3: The core architecture of the *Convolutional Neural Network* used in the inference attack.

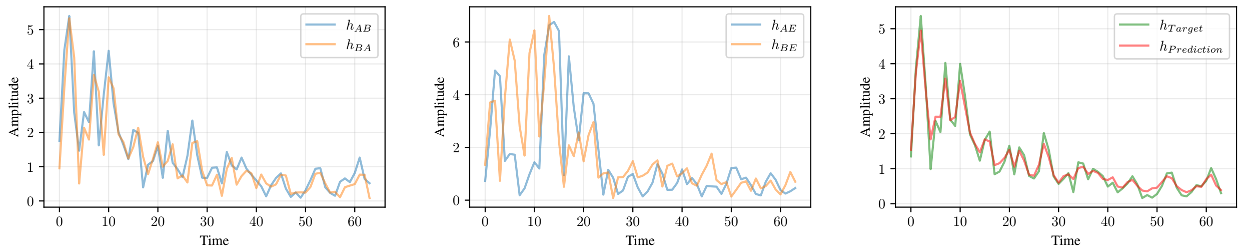
Data: We used real world UWB CIR measurements which were obtained in two separate measurement campaigns.

The first data set, dubbed *scenarios*, was recorded as described in [20]. It consists of seven different scenarios in a typical indoor office environment, ranging from entirely static to highly dynamic scenarios. The measurements itself were conducted at $4\,GHz$, with a bandwidth of $500\,MHz$ and a sampling rate of $1\,ns$.

We then conducted a second measurement campaign, resulting in a data set, which we refer to as *long-term*. Here, the three terminals (cmp. Fig. 2) were set up in a busy hallway, with Alice and Bob right across and Eve about $4\,m$ down the hall. Interferences were generated by people traversing the Line-of-Sight throughout the whole measurement. The transmission parameters are the same as for the scenarios measurement. This setup ran for about 10 hours and collected 175005 data points.

To prepare the data for the processing, the following preliminary steps were conducted for all measurements: we synchronized the CIR pairs by using the maximum of the cross correlation, in accordance to [21]. This is valid, because the attacker has all samples locally at hand, so no blind synchronization needs to be applied.

Subsequently, we purge the parts containing only noise to extract those that bear information of the CIR. To achieve this, we defined a starting point within the CIR measurement through leading edge detection. From this starting point, we took the next 64 values. Note, that values further behind the leading edge did not show significant reciprocity any more.



(a) Reciprocal measurements

(b) Measurements at Eves terminal

(c) Target values and CNN generated prediction

Fig. 4: Exemplary CIR realizations at different nodes as well as the derived target values and the values predicted by the attacking CNN.

In preparation of the training process, the CIRs were scaled by the standard deviation of the measurements.

Fig. 4a depicts the acquired measurements h_{AB} and h_{BA} after this processing. In Fig. 4b the corresponding measurements at Eve, h_{AE} and h_{BE}, are shown. The measurements of the legitimate, reciprocal channel h_{AB} and h_{BA} are called *reciprocal*, whereas Eves CIRs (overheard measurements and predictions) are called *non-reciprocal*, since they are not part of the reciprocal channel.

Finally, we defined the mean of h_{AB} and h_{BA} as target for the model training process. This decision is based in the CRKG processing itself: one of the main steps is *Information Reconciliation*, which removes slight differences between h_{AB} and h_{BA}, as those are not perfectly equal either. Hence, for an attack to succeed, it suffices to solely have the same amount or fewer differences as Alice and Bob.

Architecture: The main aim is to reconstruct the channel characteristics of the legitimate channel from the features observed at Eve. Literature indicates that *Convolutional Neural Networks* (CNN) excel at this task [12], [31]. We hence chose them when designing our core architecture. Since our input data will be one dimensional, we will use 1D convolutional layers (Conv).

The architecture of the final network used in the attack is depicted in Fig. 3. It is a straight forward CNN realization consisting of three 1DConv layers with $16, 32, 64$ feature maps respectively, followed by a 1D MaxPooling layer. Afterwards a Flatten layer reshapes the data for the subsequent processing done by two fully connected layers (Dense). The intermediate Dense layer has size 256 and the final layer has the size of the targeted output, i.e. 64 elements. After each 1DConv and before each Dense layer, a Dropout layer with rate 0.5 is applied, to avoid overfitting and to provide generalized learning results. All 1DConv and Dense Layers have Rectified Linear Activation functions.

Although this architecture might seem simple, it is sufficient to extract relevant features from the CIRs, as shown in [23]. Note, that we also tested more complex architectures like VGG, DenseNet or InceptionNet, but despite their significantly higher complexity, none of those achieved better results than the architecture presented here.

Training: To train the network, we split the acquired data by 0.7 and used the 70 % for training and the remaining 30 % for evaluation. In the case of the *scenarios* dataset, this division was made separately for each scenario. Since we intend to fit real valued output data, we used the *Mean Squared Error* as loss function and *Mean Absolute Error* as accuracy metric. Dozat suggested to train the network using the *ADAM* optimizer with Nesterov momentum for such tasks, as it can achieve the best performance [3].

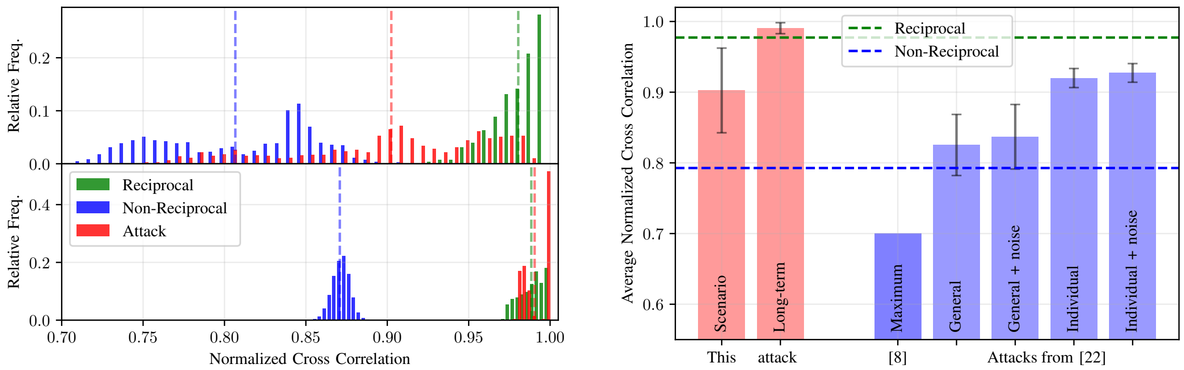
Fig. 4c shows an example of the prediction performances of the trained network.

VI. Evaluation and Discussion

To evaluate our attack, we will compare the results with those of similar attacks, i.e. [8], [22]. As both previous works evaluate their results using the maximum of the *normalized cross correlation*, we will also employ this metric, as introduced in Sec. II, to ease comparison.

Prediction accuracy: To assess the accuracy of the CNNs prediction, we first show the measurement metrics as histograms. In Fig. 5a the green and blue histogram shows the cross correlation of the reciprocal and non-reciprocal measurements, respectively. As the *scenario* data is very diverse, we expect a spread out histogram. In contrast, the *long-term* data is more stable, hence, the histograms are more consolidated.

The upper plot in Fig. 5a shows the results for the *scenario* measurements: for the reciprocal measurements the mean cross correlation is 0.980, for the non-reciprocal ones 0.807. The fact that the non-reciprocal distribution appears to be multi-modal apparently results from the different measurement scenarios: some scenarios include much interference, which translates to significantly worse results for the non-reciprocal measurements. The lower plot depicts the results for the *long-term* data. The cross correlations are more stable and higher — the non-reciprocal mean is at 0.871, the reciprocal one at 0.990, as expected.



(a) Histogram of normalized cross correlation for the data *scenarios* (upper plot) and *long-term* (lower plot) — the dotted lines represent the repsective mean.

(b) Comparison to the baseline attack, with different optimization approaches - the attack proposed here requires no optimization at all

Fig. 5: Achieved normalized cross correlation between legitimate terminals and predictions for current attack and baseline.

In the plots of Fig. 5a, the red histogram represents the cross correlation between the legitimate CIRs and the ones predicted by the CNN trained for this attack. Due to the training towards the mean of h_{AB} and h_{BA}, we expect to see a distinctively higher cross correlation than for the non-reciprocal measurements. In the *scenario* setting, the mean of attack CIRs lies at 0.902. The CIRs predicted by the attack are closer to the reciprocal than to the non-reciprocal ones, which means that the inferred CIRs substantially match the legitimate ones. Again, the histogram appears to be multi-modal due to the differing performance in the varying scenarios. For the *long-term* measurements, the average attacker correlation is 0.997 — so in fact *higher* than the average correlation of the legitimate CIRs, which is at 0.990. This is possible, as the CNN learns to predict the mean between h_{AB} and h_{BA}. Hence, a very precise prediction is closer to h_{AB} than h_{BA}. Consequently, this result means that the attacker in this setting can predict the CIRs on average so well that he has better knowledge of the reciprocal channel than the legitimate participants — and thus also of the implicit shared secret of them.

Overall, these results clearly demonstrated the high prediction accuracy of the trained attack CNN. This accuracy allows the attacker to significantly expand his information about the legitimate channel up to concrete knowledge about the channel characteristics.

Comparison to baseline attacks: In Fig. 5b we compare the prediction performance of our attack to those of the two baseline attacks in [8] and [22].

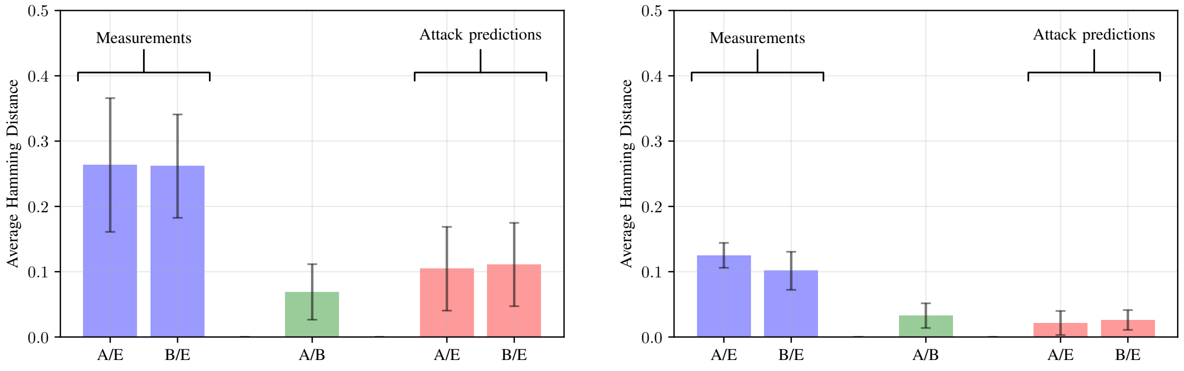
Compared to [8], our presented attack provides substantially better results for the attacker. Since only the maximum can be reliably extracted from their paper, we compare here our average value with their maximum. The attack presented here provides better cross correlation with 0.902 and 0.997 than the maximum of ≈ 0.7 in [8], despite this unfavorable comparison.

The attacks from [22] need to be differentiated regarding their optimization approach: Only results originating from the "general" optimization are comparable to our results, as only those are generalized to be applicable to all measurements in this room. With this optimization the previous work achieves average cross correlations of 0.825 and 0.837. Again, the attack presented here outperforms the previous work with a cross correlation of 0.902 and 0.997.

The individual optimizations in [22] are each specifically adapted to a concrete terminal position and room geometry. Hence, higher correlations are achieved here, with 0.920 and 0.927. However, since this specific optimization cannot be applied to other measurements but to the concrete one, these results are not comparable to the results of our generically applicable attack.

Nevertheless, in the context of the presented attack, a fair analogy to "Individual Optimization" would be an evaluation regarding the training data, since these are also specialized and not generically applicable. — here our attack would reach correlations of 0.975 and 0.999, thus again outperforming the baseline attack. However, as we are only interested in generically applicable attacks and as evaluating the training data is no meaningful analysis, these results are not shown in Fig. 5. Furthermore, our attack, even in the generic version, achieves significantly better cross correlation with 0.997 for the *long-term* measurements.

Security of keying material: The processing subsequent to the channel measurement is currently not standardized and many different solutions for the key generation exist. Hence, a definite analysis of the attack's impact on the key exchange is difficult. Nevertheless, to still show the possible implications of this attack, we implement a threshold based quantization scheme, as used in, e.g., [1], [9], [15], and analyze the resulting Hamming distances between Alice/Eve and Bob/Eve. Such threshold based quantization approaches were also used as baseline in previous quantization related work, e.g. [20], [7].



(a) *Scenario* measurements

(b) *Long time* measurements

Fig. 6: Average Hamming Distances for overheard CIRs (blue bars), legitimate CIRs (green bar) and the values predicted by the attack (red bars).

Note, that Alice and Bob would aim to achieve very low Hamming distance between each other, and as high Hamming distances as possible to the sequence Eve infers.

As we trained our CNN to predict the mean of h_{AB} and h_{BA}, the resulting quantized bit vector of the predicted CIRs are expected to have significantly lower Hamming distances than quantized h_{AB}/h_{AE} or h_{AB}/h_{BE}.

Fig. 6a presents the Hamming distances for the observations obtained by Eve, Alice and Bob in the *scenario* data set as well as for the predictions generated by the presented attack. The reference is the green bar in the middle showing the average Hamming distance of 0.069 for the quantized CIRs of Alice and Bob. The two blue bars show the average Hamming distance between the overheard CIR measurements at Eve and Alices/Bobs CIR after quantization, with 0.263 and 0.261, respectively. On the right, the red bars show the Hamming distance between the attack predictions and Alices/Bobs quantized CIRs, achieving 0.105 and 0.110, respectively. It is visible, that the attack generates values well within the standard deviation of the legitimate channel, i.e. binary sequences which should be successfully corrected during *Information Reconciliation*. Further, the attack yielded a perfect match, i.e. a Hamming distance of 0, in 1.2% of all cases and Hamming distances below those of Alice and Bob in 39.1% of all cases.

In combination, this means that an adversary carrying out this attack can derive the same key bits as the legitimate communication partners in at least 39.1% of all cases.

Fig. 6b depicts the same values as measured for the *long-term* data set. Since these measurements, unlike the previous ones, do not contain highly dynamic interference, all distances are considerably lower. The green Alice/Bob reference is 0.032; the respective results for Alice/Eve and Bob/Eve are 0.125 and 0.101. In this setting, the CNN attack achieved significantly better results: with average Hamming distances of 0.021 and 0.026, the predictions achieved even better results, than the legitimate Alice/Bob CIR pairs. This is possible, because we trained the network to predict the mean of Alice and Bobs measurements. Hence, if the attack prediction is accurate enough, the resulting Hamming distance at Eve can even be lower than those of Alice and Bob. Further, the attack yields Hamming distances of 0 in 33.5% of all cases and distances below the legitimate on in 83.5% of all cases.

Again, this means that the attack can derive the correct key bits in 83.5% of all cases.

In conclusion, these results show that a significant proportion of the key material can be successfully inferred, using the proposed attack. Within both datasets, the CIRs predicted by the attack show high correlation to the legitimate CIRs and in many cases even exceed the correlation of the reciprocal measurements, resulting in a significant proportion of the keying material being compromised.

Using RSSI instead of CIR-based schemes cannot improve security, as indicated above. To the contrary, since RSSI can directly be derived from CIR, RSSI-based systems will be much more vulnerable to our attack.

Considering that other PhySec primitives have different input data requirements, the results of the *long-term* measurement reveal a further problem: since these measurements were recorded with comparatively little dynamic range in the channel, they would also be suitable for PhySec primitives requiring more stable channel characteristics, such as authentication. The obtained results of the presented attack imply that the underlying assumption of uncorrelated observations is not unconditionally acceptable. Hence, similar PhySec primitives relying on the same core assumption may be affected in the same way as the presented CRKG.

VII. Summary and Outlook

In this paper we have investigated how one of the most fundamental assumptions of Physical Layer Security — that adversaries residing further than half a wavelength away from the legitimate parties can only overhear uncorrelated measurements — can successfully be attacked.

To achieve this, we use the fact that core properties of the environment can be derived from monitored channel impulse responses. With the help of these, it is in turn possible to infer the channel properties between the legitimate terminals. For that purpose, we propose an attack based on Convolutional Neural Networks, which learns to directly infer seemingly secret CIRs from CIRs that are observed at a third, remote location.

We thoroughly evaluated this attack using two sets of real world measurements — one consisting of different typical indoor scenarios, and another indoor long-term measurement. Our attack infers the legitimate CIRs with very high accuracy in both sets: in the first more dynamic scenario, the inferred CIRs reach average cross correlations of 0.902 compared to the correlation of the legitimate ones, 0.980. In a data set of long-term measurements, the predictions even outperformed the CIR estimation between the legitimate parties, with correlations of 0.997 vs. 0.990, due to specific training.

We also analysed the Hamming distances between the quantized legitimate and inferred CIRs, to understand the implications for Channel Reciprocity-Based Key Generation (CRKG). The results clearly show that a significant proportion of the potential key material must be considered predictable, and hence insecure. More precisely, 39.1% of the key material is compromised in our first measurements, and 83.5% in the case of long-term measurements.

Our investigations demonstrate that the presented attack poses a significant threat to PhySec primitives. This apparent risk does not only concern the presented use case of CRKG, but also other PhySec methods based on CSI in the form of CIRs, such as authentication.

To improve the attack results, especially for non-static scenarios, we are currently in the process of recording further CIR measurements including the location information within the room. We will use this additional input in the training process of the CNN and expect to achieve higher prediction accuracy for dynamic settings. Furthermore, we will investigate how such attacks can be defended or mitigated in the PhySec context, e.g. by additional preprocessing or by applying machine learning on the defender side, as well.

References

[1] T. Aono *et al.*, "Wireless secret key generation exploiting the reactance-domain scalar response of multipath fading channels: RSSI interleaving scheme," in *Wireless Technology*, 2005.

[2] N. Döttling, D. E. Lazich, J. Müller-Quade, and A. S. de Almeida, "Vulnerabilities of wireless key exchange based on channel reciprocity," in *WISA*, ser. Lecture Notes in Computer Science. Springer, 2010.

[3] T. Dozat, "Incorporating nesterov momentum into adam," 2016.

[4] S. Eberz, M. Strohmeier, M. Wilhelm, and I. Martinovic, "A practical man-in-the-middle attack on signal-based key generation protocols," in *ESORICS*, ser. Lecture Notes in Computer Science. Springer, 2012.

[5] M. Edman, A. Kiayias, and B. Yener, "On passive inference attacks against physical-layer key extraction?" in *Proceedings of the Fourth European Workshop on System Security*, 2011.

[7] R. Guillaume *et al.*, "Fair comparison and evaluation of quantization schemes for phy-based key generation," in *Int. OFDM Workshop*, 2014.

[6] A. Goldsmith, *Wireless Communications*. Cambridge Univ. Press, 2005.

[8] S. T.-B. Hamida, J.-B. Pierrot, B. Denis, C. Castelluccia, and B. Uguen, "On the security of uwb secret key generation methods against deterministic channel prediction attacks," in *IEEE Vehicular Technology Conference (VTC Fall)*, 2012.

[9] S. Jana *et al.*, "On the effectiveness of secret key extraction from wireless signal strength in real environments," in *International Conference on Mobile computing and networking*, 2009, pp. 321–332.

[10] R. Jin and K. Zeng, "Physical layer key agreement under signal injection attacks," in *IEEE Conference on Communications and Network Security (CNS)*, 2015.

[11] J. Kunisch and J. Pamp, "Radio channel model for indoor UWB WPAN environments," IEEE 802.15.3a, Tech. Rep. IEEE P802.15-02/281, 2002.

[12] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems*, 2009.

[13] F. J. Liu, X. Wang, and H. Tang, "Robust physical layer authentication using inherent properties of channel impulse response," in *Military Communications Conference*, 2011.

[14] H. Liu, Y. Wang, J. Yang, and Y. Chen, "Fast and practical secret key extraction by exploiting channel response," in *IEEE INFOCOM*, 2013.

[15] S. Mathur *et al.*, "Radio-telepathy: Extracting a secret key from an unauthenticated wireless channel," in *Mobile Computing and Networking*. ACM, 2008.

[16] S. Mitra, *Signals and Systems*, ser. Oxford Series in Electrical and Computer Engineering. Oxford University Press, 2016.

[17] A. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE JSAC*, 1987.

[18] W. Trappe, "The challenges facing physical layer security," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 16–20, 2015.

[19] J. K. Tugnait and H. Kim, "A channel-based hypothesis testing approach to enhance user authentication in wireless networks," in *International Conference COMmunication Systems NETworks (COMSNETS)*, 2010.

[20] P. Walther *et al.*, "Improving quantization for channel reciprocity based key generation," in *IEEE LCN*, 2018.

[21] P. Walther, E. Franz, and T. Strufe, "Blind synchronization of channel impulse responses for channel reciprocity based key generation," in *IEEE LCN*, 2019.

[22] P. Walther, R. Knauer, and T. Strufe, "Passive Angriffe auf kanalbasierten Schlüsselaustausch," in *SICHERHEIT 2020*, D. Reinhardt, H. Langweg, B. C. Witt, and M. Fischer, Eds. Bonn: Gesellschaft für Informatik e.V., 2020, pp. 41–51.

[23] P. Walther and T. Strufe, "Blind twins: Siamese networks for Non-Interactive information reconciliation," in *International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2020.

[24] C. Wan, L. Wang, and V. V. Phoha, "A survey on gait recognition," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–35, 2018.

[25] L. Xiao, L. J. Greenstein, N. B. Mandayam, and W. Trappe, "Using the physical layer for wireless authentication in time-variant channels," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, 2008.

[26] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, 2013.

[27] M. Zafer, D. Agrawal, and M. Srivatsa, "Limitations of generating a secret key using wireless fading under active adversary," *IEEE/ACM Trans. Netw.*, vol. 20, 2012.

[28] C. Zenger, "Physical-layer security for the internet of things." Ph.D. dissertation, Ruhr University Bochum, Germany, 2017.

[29] C. Zenger, H. Vogt, J. Zimmer, A. Sezgin, and C. Paar, "The passive eavesdropper affects my channel: Secret-key rates under real-world conditions," in *IEEE Globecom Workshops*, 2016.

[30] J. Zhang, R. Woods, T. Q. Duong, A. Marshall, Y. Ding, Y. Huang, and Q. Xu, "Experimental study on key generation for physical layer security in wireless communications," *IEEE Access*, 2016.

[31] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017.