Detecting Relay Attacks with Timing-Based Protocols

Jason Reid*, Juan M. González Nieto*, Tee Tang+, Bouchra Senadji+

*Information Security Institute, + School of Engineering Systems
Queensland University of Technology
{jf.reid, j.gonzaleznieto, t.tang, b.senadji}@qut.edu.au

Abstract. Distance bounding protocols have been proposed as means of detecting relay attacks, also known as *mafia fraud*. In this paper we present the first symmetric key based distance bounding protocol that is also resistant to so-called *terrorist fraud*, a variant of mafia fraud. Relay attacks present a serious threat to RF security devices (contactless smart cards, RFID tags and the like) because they undermine the implicit assumption that the device is physically close to the reader when it is operating. In applications such as physical access control this assumption of physical proximity is all-important. Distance bounding protocols require a communication channel that can exchange single bits with extremely low latency. This unconventional communication requirement has prompted Hancke and Kuhn to assert in a recent publication [12] that ultra wide band (UWB) radio is necessary to achieve a useful distance bounding resolution. We analyse this assertion and present a alternative, novel communication approach that leverages the phenomena of side channel leakage to deliver a low latency channel. Our proposal is capable of detecting sophisticated relay attacks without resorting to the considerable expense and complexity of UWB radio. We present experimental results to support our arguments.

1 Introduction

Recent publications [15, 11] have highlighted the vulnerability of RF transponders to relay attacks by presenting practical, low cost attacks on ISO 14443 contactless smart cards. The relay attack is particularly insidious because it works without the need to circumvent any cryptographic security protocols that may be in place. ISO 14443 cards have a short operating range of 10 cm from the card reader [1]. There is an implicit assumption that if a reader is communicating with a card, that card is actually within 10 cm of the reader. However this assumption may not be well founded because an attacker can simply relay the RF messages from the reader to a legitimate card that is far away and relay the card's responses back to the reader.

Distance bounding protocols have been proposed as a means of protecting against relay attacks. A distance-bounding protocol is an authentication protocol between a *prover* A and a *verifier* B, whereby B obtains corroborating evidence about A's claimed identity and physical proximity at the time the protocol is run. Distance-bounding protocols can be thought of as traditional identification protocols enhanced with a distance-bounding mechanism. The former provides assurance as to the identity of the prover, while the latter allows the verifier to upper bound the distance which separates her from the prover. This dichotomy of distance-bounding protocols into an identification mechanism and a distance-bounding mechanism is readily seen in most proposals, which can be easily decomposed into an identification part, matching some known identification protocol, and a distance-bounding part, consisting of multiple fast challenge-response rounds. The distance between prover and verifier is upper-bounded by measuring the time intervals between challenges being sent and responses being received.

$$A \longleftrightarrow \bar{B} \longleftrightarrow\!\!\!-\!\!\!-\!\!\!-\!\!\!-\!\!\!-\!\!\!\longrightarrow \bar{A} \longleftrightarrow B$$

Fig. 1. Adversarial setting

The first distance-bounding protocol was proposed by Brands and Chaum [4] to thwart *mafia fraud* - the name for relay attacks against identification protocols first described by Desmedt [8]. In a mafia fraud, the adversary consists of two parts: a rogue prover $\bar{A}$ and a rogue verifier $\bar{B}$, sitting in between the real verifier B and prover A as shown in Figure 1. $\bar{A}$ and $\bar{B}$ simply relay the protocol messages between A and B. Hence what the adversary achieves is to fool B into thinking that he is directly communicating with A, when in reality he is talking to $\bar{A}$. This attack does not violate the traditional security requirements of identification protocols, however it may be a concern if the verifier incorrectly makes assumptions as to the proximity of the prover. For example, consider the case where B is an RF reader enforcing access control through a door and A uses a RF proximity card to authenticate to B. A succesful mafia fraud attack would allow an adversary to open the door when A is sitting at a restaurant close by to $\bar{B}$ who is stealthily running the identification protocol with A's card and relaying all the information to $\bar{A}$, who is present near the door running the identification protocol with B using the messages received from her accomplice. Brands and Chaum [4] described two protocols secure against mafia fraud attacks. The underlying identification protocols are a signature based challenge-response mechanism for one of them, and a zero-knowledge identification protocol for the other. These protocols use public key cryptographic operations, which are computationally demanding for highly resource constrained devices such as RFID tags.

1.1 Contribution and overview

The paper is presented in two parts, the first dealing with protocol specifics and the second with implementation. Recently, Hancke and Kuhn [12] proposed a very efficient distance-bounding protocol which is secure against mafia fraud. In Section 2 we review their protocol and explain why it is not resistant to *terrorist fraud*. In Section 3 we propose the first symmetric key based distance-bounding protocol which is resistant to terrorist fraud and is computationally efficient enough to be implemented in resource constrained devices.

Hancke and Kuhn [12] have proposed the use of ultra wide band radio (UWB) to meet the demanding communications requirements of their distance bounding protocol. The addition of UWB would add appreciable cost and complexity to contactless smart card integrated circuits, so it could only be justified in the absence of simpler, lower cost alternatives. In Section 4 we present an analysis of the communication channel requirements for distance bounding. This analysis highlights the importance of low communication latency, which is not purely a function of the channel bit rate. We propose a novel communication method based on the principle of *side channel leakage*, that has very low latency. In Section 5 we report on our current investigations into adapting the existing load modulation circuitry in proximity cards to our proposed communication technique. We present experimental results indicating that a modified load modulation scheme can provide sufficient timing resolution to detect sophisticated relay attacks launched by well funded attackers. Our proposed approach avoids the additional cost and complexity of adopting UWB radio.

The following notation is used in the rest of the paper.

- We use ← to indicate assigment to a variable. If A is a set the $x \leftarrow A$ assigns to x a random element of X according to the uniform distribution.
- $\{0,1\}^n$ denotes the set of all strings of bit-length n.
- Given a string s, we use s_i to denote the i^{th} less significant bit of s;
- ID_U is the identity string corresponding to user U.
- *time*() is a function implemented at all parties that returns the internal clock time. To measure the time between two events we use two instructions, `Start clock` and `Stop clock`, such that `Start clock` Δt assigns $t_o = time()$ and `Stop clock` Δt, assigns $t_f = time()$ and $\Delta t = t_f - t_o$. Note that we do not require clocks at different parties to be synchronised.

2 Hancke and Kuhn's distance bounding protocol

Hancke and Kuhn's [12] distance-bounding protocol is highly efficient. The protocol (see Figure 1) is based on a symmetric-key identification mechanism, where the prover and verifier share a common secret value s. The distance is parametised by the maximum challenge-response delay allowed, $\Delta t_{\max}$. The protocol starts by having A and B exchange random nonces r_A and r_B. The prover then applies a keyed hash function[1] h to the concatenation of the nonces $r_A \| r_B$ to get

[1] Hancke and Kuhn [12] specifically require that this function be one-way and collision resistance. We notice however that this is clearly insufficient and that h should be pseudo-random. Otherwise, if not all the bits in d are pseudo-random, an adversary will have an advantage in responding to a proportion of the challenges. Also note that pseudo-randomness implies collision resistance and one-wayness.

d. The prover splits d into two n-bit strings l and r. A fast n-round challenge-response phase begins then. At each round, B sends challenge bit α_i, to which A must respond with the i^{th} bit of l if $\beta_i = 0$, and the i^{th} bit of r if $\beta_i = 1$. The verifier checks that the received response is correct. (He can do so, since he can also compute l and r.) Additionally, B measures the time Δt_i elapsed between challenge and response. B makes sure that all delays Δt_i are less than the bound $\Delta t_{\max}$. If all checks are succesful, B outputs `accept`, otherwise B outputs `reject`.

Shared Information: Secret key s.

A		B
(Prover)		(Verifier)
$r_A \leftarrow \{0,1\}^m$	$\xleftarrow{\qquad r_B \qquad}$	$r_B \leftarrow \{0,1\}^m$
$d \leftarrow h(s, r_A \Vert r_B)$	$\xrightarrow{\qquad r_A \qquad}$	$d \leftarrow h(s, r_A \Vert r_B)$
$l \leftarrow \beta_1 \Vert \cdots \Vert \beta_n$		$l \leftarrow \beta_1 \Vert \cdots \Vert \beta_n$
$r \leftarrow \beta_{n+1} \Vert \cdots \Vert \beta_{2n}$		$r \leftarrow \beta_{n+1} \Vert \cdots \Vert \beta_{2n}$
		$\alpha \leftarrow \{0,1\}^n$
	`For` $i = 1$ `to` n `do:`	
	$\xleftarrow{\qquad \alpha_i \qquad}$	`Start clock` Δt_i
$\beta_i \leftarrow \begin{cases} l_i & : \alpha_i = 0 \\ r_i & : \alpha_i = 1 \end{cases}$	$\xrightarrow{\qquad \beta_i \qquad}$	`Stop clock` Δt_i `Check` β_i `Check` $\Delta t_{\max} \geq \Delta t_i$
	`End for`	

Protocol 1: Distance bounding protocol resistant against terrorist attacks

If B accepts, assuming that information cannot travel at a speed higher than the speed of light c, then the distance between A and B is upper-bounded [2] by $c\Delta t_{\max}/2$.

Hancke and Kuhn [12] showed that the probability that a mafia fraud attacker has to make B falsely accept is bounded by $\left(\frac{3}{4}\right)^n$.

Terrorist fraud Desmedt [8] considered another type of active attacks against identification protocols, which he called *terrorist attacks*. Here, unlike mafia fraud attacks where the prover is oblivious to the attack that is underway, the prover conspires with $\bar{A}$ and $\bar{B}$ to intentionally try to fool the verifier as to A's location. Defending against terrorist attacks is more difficult, since A's secret information

[2] A better bound can be obtained when we know the time Δt_p that it takes for A to process a challege. In this case, the distance is bounded by $c(\Delta t_{\max} - \Delta t_p)/2$.

(e.g. authentication keys) may be used in a manner which is differerent to what the protocol prescribes. Clearly, if A is prepared to released her secret authentication keys to $\bar{A}$, then the attack is trivially successful. When dealing with terrorist attacks, we preoccupied ourselves with attacks where the prover does not reveal to her accomplices secret information that will allow the accomplices to impersonate A in more than a single run of the protocol. In particular, A does not reveal her long term private key.

Protocol 1 is not secure against terrorist fraud attacks. A remote A can always relay r and l to a rogue prover $\bar{A}$ who is close by to B. Note that the time-critical phase does not start until B sends the first challenge bit, and that releasing r and l does not compromise the long-term secret s.

To the best of our knowledge, the only distance-bounding protocol that protects against terrorist fraud attacks is the protocol of Bussard [5]. His solution is public-key based and uses zero-knowlege techniques, which makes it computationally expensive, especially for implementation in low-cost RF computing devices, such as RFID tags. The basic idea of Bussard's protocol is to force the prover to give away her private key in order to mount a terrorist attack. The prover computes $c = \mathcal{E}_k(sk_A)$, the encryption of her long-term (important) private key sk_A under a newly generated session key k. The verifier then sends challenge bits α_i to the prover. If $\alpha_i = 1$, the prover must respond with the bit c_i from the ciphertext. If $\alpha_i = 0$, the prover returns the bit k_i of the session key. Thus, in order to successfully and timely reply to the challenges the prover must be in possesion of c and k, and therefore of sk_A. This is better illustrated in Section 3, where we apply the same basic idea to Protocol 1.

Distance Fraud Finally, there is one more attack against distance-bounding protocols that must be considered, and which Bussard [5] refers to as *distance fraud*. In this type of attack the prover on her own tries to subvert the security of the protocol by having the verifier believe she is close by, when in fact she is not. This is the simplest attack that distance-bounding protocols should defend against. Note that terrorist fraud resistance implies distance fraud resistance.

3 New distance-bounding protocol

As indicated in Section 2, Bussard's protocol [5] protects against terrorist fraud attacks, but its use of asymmetric techniques makes it computationally demanding. On the other hand, the more computationally efficient protocols published, based on symmetric key authentication, do not afford terrorist fraud resistance. In this section, we propose the first symmetric key based distance-bounding protocol which is resistant to terrorist fraud attacks and is efficient enough for implementation in low cost devices. We enhance the protocol of Hancke and Kuhn [12], which is to our knowledge the most efficient mafia-fraud resistant protocol, by applying the basic idea behind the terrorist fraud resistance of Bussard's protocol. The result is shown as Protocol 2. The efficiency of the new protocol remains practically unchanged with respect to Hancke and Kuhn's [12], the main difference being the addition of a symmetric encryption (in practice, an xor operation as discussed below).

Shared Information: Secret key s.

$$\begin{array}{ccc}
A & & B \\
\text{(Prover)} & & \text{(Verifier)} \\
r_A \leftarrow \{0,1\}^m & \xleftarrow{\qquad ID_B, r_B \qquad} & r_B \leftarrow \{0,1\}^m \\
k \leftarrow \mathrm{KDF}(s, ID_A \| ID_B \| r_A \| r_B) & & \\
c \leftarrow \mathcal{E}_k(s) & \xrightarrow{\qquad ID_A, r_A \qquad} & k \leftarrow \mathrm{KDF}(s, ID_A \| ID_B \| r_A \| r_B) \\
 & & c \leftarrow \mathcal{E}_k(s), \alpha \leftarrow \{0,1\}^n \\
 & \texttt{For } i = 1 \texttt{ to } n \texttt{ do:} & \\
 & \xleftarrow{\qquad \alpha_i \qquad} & \texttt{Start clock } \mathit{\Delta} t_i \\
\beta_i \leftarrow \begin{cases} c_i & : \quad \alpha_i = 0 \\ k_i & : \quad \alpha_i = 1 \end{cases} & \xrightarrow{\qquad \beta_i \qquad} & \begin{array}{l} \texttt{Stop clock } \mathit{\Delta} t_i \\ \texttt{Check } \beta_i \end{array} \\
 & \texttt{End for} &
\end{array}$$

Protocol 2: Distance bounding protocol resistant against terrorist attacks

In Protocol 2, we have made explicit the identities of prover and verifier by adding them in the initial exchange of nonces. This is considered sound protocol engineering practice. Both A and B now use a key derivation function KDF to derive a symmetric encryption key k, which is used to encrypt the long-term shared secret s. The range of both KDF and $\mathcal{E}$ is $\{0,1\}^n$. The fast challenge-response phase of the protocol is similar to Hancke and Kuhn's, except that now the i^{th} bit of the ciphertext c is returned when $\alpha_i = 0$ and the i^{rm} bit of the key k otherwise. Note that knowledge of k and c is equivalent to knowing the shared secret s, since $s = \mathcal{D}_k(c)$.

Security We will now show that Protocol 2 is secure against distance, mafia and terrorist fraud. We do so by proving that the success probability of any realistic (i.e. polynomial-time) adversary is negligible[3] in n. The assumptions made in the following analysis are as follows:

1. KDF is pseudo-random, i.e. when s is a secret of high enough computational entropy, the output of the function is computationally indistinguishable from uniformly random. In practice, KDF can be a MAC algorithm such HMAC [3].

[3] A probability function $\epsilon(n)$ is said to be negligible on n when, asymptotically, it grows slower than the inverse of all polynomials in n.

2. $\mathcal{E}$ is a semantically secure encryption function, i.e. an adversay does not learn any (computational) information about the plaintext. In practice, because the strings to be encrypted are short and the key varies for each run of the protocol, we can use a one-time pad, i.e $\mathcal{E}_k(s) = s \oplus k$.

Theorem 1. *Protocol 2 is secure against distance fraud.*

Proof. To mount a distance fraud attack, the prover, who is not close to the verifier, must respond to the challenges within the short interval $\Delta t_{\max}$. Since A is not close by, she must send the response before she receives the challenge. The best she can do is to guess the challenge and answer according to that guess. This guess will be correct with a probability of $1/2$, and hence the probability that B accepts is $(1/2)^n$, which is negligible. □

Theorem 2. *Protocol 2 is secure against mafia fraud.*

Proof. (Sketch) Here the adversarial setting corresponds with the one depicted in Figure 1. A is honest, in the sense that she does not cooperate with the attackers $\bar{A}$ and $\bar{B}$, and A is not close to B, which implies that it is not physically possible for $\bar{A}$ and $\bar{B}$ to pass on the challenge to A, get the response from A and relay it back to B in time. k is different for each run of the protocol with overwhelming probability due to the inclusion of the nonce r_A in the key derivation function. This together with assumptions 1 and 2 above, implies that it is impossible for any adversary to guess any bit k_i or c_i with probability non-negligiby different from $1/2$. Hence the best $\bar{A}$ and $\bar{B}$ can do is guessing the challenge bit before it is output by B and send it to A. This could be done before the challenge response phase starts. For example, $\bar{B}$ withholds message 2 for a time long enough to allow him to complete a run of the protocol with A, using challenges $\bar{\alpha}_i$ chosen by $\bar{B}$ himself. $\bar{B}$ then passes to $\bar{A}$ the value r_A, the challenges $\bar{\alpha}_1, \ldots, \bar{\alpha}_n$, and the responses $\bar{\beta}_1, \ldots, \bar{\beta}_n$. $\bar{A}$ then completes the protocol with B. Since B choses the challenges α_i uniformly at random, on average only half of the challenges $\bar{\alpha}_i$ will coincide. When this happens, then $\bar{A}$ can send the valid response $\bar{\beta}_i$; otherwise, $\bar{A}$ can only guess the right reponse with probability $1/2$. Overall, the probability that $\bar{A}$ and $\bar{B}$ fool the verifier into accepting is essentially $(3/4)^n$, which again is negligible. □

Theorem 3. *Protocol 2 is secure against terrorist fraud.*

Proof. (Sketch) It follows from Theorem 2 that in order to have B accepting, someone close to B must have k and c, which only A can generate. Since knowing k and c implies knowing $s = \mathcal{D}_k(c)$, and since A is assumed not to give away her long-term secret s, we must conclude that it is A who is close by. □

Noise errors In practice, as discussed by Hancke and Kuhn [12] and further elaborated in Section 4, the communications link between prover and verifier during the fast challenge-response phase is unreliable. This means that the protocol should tolerate transmission errors during that phase, by increasing the number of challenge-reponse rounds according to the expected error rate. We refer the reader to Hancke and Kuhn's paper [12] for the quantitative analysis.

Comparison with existing schemes Table 1 presents the list of distance-bounding protocols that we have found in the literature and classifies them according to the following criteria:

- *Identification technique:* whether the underlying identification protocol is based on asymmetric, symmetric or zero-knowledge techniques (see Menezes [17]).
- *Unilateral/Mutual:* whether the protocol authenticates the identity and proximity of one or both participants.
- *Mafia fraud resistance:* does the protocol defend against mafia fraud attacks.
- *Terrorist fraud resistance:* does the protocol defend against terrorist fraud attacks.
- *Distance fraud resistance:* does the protocol defend against distance fraud attacks.

The new proposal is the only scheme that employs symmetric authentication techniques and is secure against all types of attacks. The protocol by Capkun *et al.* [6] provides mutual entity authentication as well as proximity authentication, i.e. A and B are both prover and verifier of each other. We have tried to design a mutual distance-based protocol version of the new protocol, however, in terms of efficiency, it does not appear possible to do significantly better than running the fast challenge-response phase twice.

Protocol	Type	Uni/Mut	Dist.	Mafia	Terr.
Brands & Chaum-I [4]	asym	uni	yes	yes	no
Brands & Chaum-II [4]	zk	uni	yes	yes	no
Sastry *et al.* [18]	sym	uni	yes	no	no
Capkun *et al.* [6]	sym	mut	yes	yes	no
Capkun & Hubaux [7]	sym	uni	yes	yes	no
Hancke & Kuhn [12]	sym	uni	yes	yes	no
Bussard [5]	zk	uni	yes	yes	yes
New proposal	sym	uni	yes	yes	yes

Table 1. Comparison of existing distance-bounding protocols

4 Communications requirements for to distance bounding

In this section we analyse requirements and implementation issues associated with the communications channel used in the time critical phase. We propose a novel communication approach that exploits the underlying principle of a security vulnerability, namely *side channel leakage*, in a constructive way to provide the necessary distance bounding resolution for constrained devices (particularly, ISO 14443 contactless smart cards).

The communication requirements of a distance bounding protocol are both demanding and unconventional - to achieve useful distance resolution they require extremely low communication latency but they do not require a correspondingly high bit rate since in the time critical phase, they exchange single bits punctuated by processing delay. A communication channel that can detect and correct errors may actually be a disadvantage because reliability mechanisms introduce overheads; more bits need to be exchanged but more importantly, an additional and possibly variable number of processing cycles is required to detect and correct errors. As we noted in Section 3, bit errors caused by channel noise can be tolerated by simply by increasing the number of challenge response rounds.

The reason why variable processing time for the prover is a problem is because it makes it difficult to isolate signal propagation time, which is the goal of the time critical phase of a distance bounding protocol. Total elapsed round trip time for a challenge response round comprises two components: processing time and propagation time. To reliably isolate the propagation component, the processing component should be small and invariant. For inductively powered devices, propagation time is going to be a very small percentage of total round trip time. For example, an ISO 14443 contactless smart card has a maximum operating distance of 10 cm [1] so the round trip signal propagation time is two thirds of a nanosecond. At the standard reader supplied clock rate of 13.56MHz, a single clock cycle takes 74 ns, enough time for a signal to propagate 22 m. Variation in processing time presents a real challenge to the accuracy of timing based solutions. An attacker may be able to accelerate a legitimate card's processing by providing it with a higher frequency clock signal (overclocking) to absorb the delay introduced by a relay attack.

4.1 Preventing accelerated prover processing

ISO 14443 type contactless smart cards are passive devices that receive their power via inductive coupling with the 13.56 MHz magnetic alternating field generated by a reader device's antenna coil [1]. The standard requires the reader to supply the RF operating field within a tolerance of ±7 kHz. Therefore, where a card uses the operating field as the source of its internal clock signal, (as is common) it only needs to accept a frequency that is 0.05% greater than 13.56 MHz. There are two main approaches to stop an attacker from operating a card at a higher than intended frequency; phase locked loop (PLL) internal clock generators and high frequency filters. Internal PLL-based clock signal generators are an increasingly popular choice, particularly in microprocessor cards. They have the advantage that the frequency of the generated signal is independent of the reader-supplied frequency. In the second case where the card uses a reader-supplied clock signal, high frequency protection usually takes the form of a low pass filter which resets the card when the filter threshold is exceeded. Tolerances of the order of a few percent are possible. It is therefore reasonable to assume that for appropriately designed hardware, an attacker can be limited to overclocking by no more than a few percent, thereby absorbing only 2-3 ns of introduced delay per clock cycle of calculation.

We also assume that the attacker cannot economically defeat a legitimate card's protection mechanisms to extract the long term secret key to transfer it to a faster device. This allows us to conclude that any successful distance bounding protocol run was executed with the real card. Under these assumptions the amount of introduced delay that can be absorbed is largely determined by the prover's clock frequency and the number of clock cycles required to compute the response, (ignoring implementation specifics of the communication method which we examine shortly). If the number can be kept small, the verifier can account for the processing time with some accuracy, thus permitting a reliable allocation of the portion of total round trip time to propagation.

4.2 A new approach to low latency communication

Keeping the number of clock cycles small within the existing communication architecture of contactless smart cards is not possible because after initialisation, they communicate via a reliable layered transport protocol. It is impractical to reliably detect relay attacks on these devices using their existing communication protocols, in part because of the large number of clock cycles required to process the communication - there is too much opportunity to accelerate processing to absorb introduced propagation delay.

To address this problem, we propose a new approach to communication that addresses the unconventional requirements of distance bounding protocols. The essential element of our proposal is that the verifier measures a physical side effect of the calculation process and from this, infers the result. The general principle that underlies this approach is not new. However, up until now, communications channels of this type have been thought of as a serious security vulnerability, certainly not purposely optimised and used to achieve legitimate protocol goals. So in this sense, the proposal is to the best of our knowledge, novel. The following sections describe the technique in the context of ISO 14443 contactless smart cards, though it is worth noting that it is also applicable to contact smart cards and a range of other devices that are also vulnerable to relay attacks (e.g. see [10] for a description of relay attacks on contact smart cards).

The proposed communication technique exploits the principle of simple side-channel analysis (SSCA) to disclose the response bit to the verifier in the time critical phase. SSCA is based on the following observation: computation in a microprocessor is the result of the physical process of electrons moving through semiconductor gates. This physical process takes a finite amount of time and consumes a measurable amount of energy which is radiated as heat and electromagnetic emanations. Such measurable phenomena are known as *side channel leakage*. Careful analysis of these phenomena can reveal detailed information about the internal state of the device [16].

The key advantage of communicating through side channel leakage is the dramatic reduction in latency - the verifier can detect the response bit as the prover calculates it, effectively being able to watch the prover in real time as it ‘thinks’. This eliminates the variable delays that would be introduced if the response bit had to traverse a layered communications stack before the verifier received it. While both power and EM side channels are possible candidates, power has advantages in implementation simplicity. Since the verifier provides the card’s power it is ideally placed to precisely monitor it.

What modifications are required to implement a usable power communication side channel? The card needs to implement a special XOR instruction that is designed to have significantly different power consumption characteristics, depending on the value of the output bit. This can be achieved by having the instruction conditionally switch a resistor into the supply circuit only if the result is for example, binary one. Normally, simple side channel analysis requires a digital oscilloscope with a reasonably high sampling frequency to analyse the power consumption since the data dependent behaviour is subtle. This is not necessary with the proposed method because the special XOR instruction can be implemented in such a way that it produces a deliberately large power consumption spike that can be detected with a simple peak detector circuit in the reader. Moreover, within the time critical phase, spikes of this magnitude would only be produced by the special instruction. The signaling load should be conditionally switched, ideally in the same clock cycle that the XOR result is calculated. The essence of the proposal is that the switching should be as integral to the instruction as the calculation itself. The circuitry required to implement this instruction and the corresponding peak detector circuit can be implemented at modest cost.

4.3 Communication latency and distance bounding resolution

Realistic attack scenarios on contactless smart cards may involve relay distances as small as a few meters since legitimate cards are often found in the close vicinity of a card reader. The propagation delay for this distance is of the order of 10-20 ns, so at first glance it would appear that a useful distance bounding protocol would need to detect delays of 10-20 ns. This is a technically demanding requirement. ISO 14443 contactless smart cards support a base communication rate of 106 kbits/s. At this bit rate a signal propagates 2.8 km in one bit period. Hancke [12] argues that this bit rate is inadequate for distance bounding and that contactless smart cards require Ultra Wide Band (UWB) radio to achieve the necessary distance resolution to detect such short range attacks. Since the addition of UWB radio to smart card integrated circuits would add appreciable cost and complexity to a very cost sensitive product, it is worth examining whether short range attacks can be detected using existing technology, applied according to our proposed side channel leakage communication model.

ISO 14443 contactless smart cards communicate with the reader via load modulation. A resistor in the card’s power supply circuit is switched in and out of circuit in time with the data to be transmitted. When the resistor is switched in, the card consumes more power and this increased consumption can be sensed as an amplitude change (as measured in the reader’s antenna circuit) in the 13.56 MHz carrier (f_c). For Type A cards, the data is baseband encoded

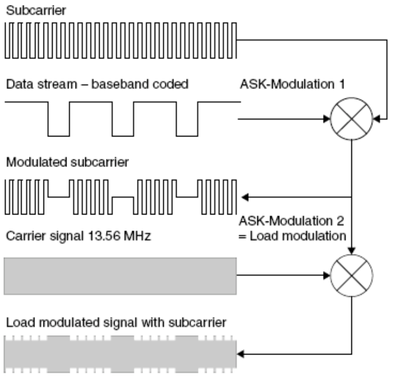

Fig. 2. Load modulation with a modulated subcarrier [9].

using Manchester binary coding and this baseband signal is modulated on a 847 kHz subcarrier using amplitude shift keying (ASK). This modulated subcarrier signal is used to turn the load resistor on and off to modulate f_c (see Figure 2). Subcarrier modulation is used instead of direct baseband data modulation on f_c because of associated benefits with the frequency spectrum it produces. The subcarrier generates two sidebands that contain the coded data. The sidebands appear 847 kHz either side of f_c . Since they are a reasonable spectral distance away from the main carrier frequency, a simple filter in the reader's receive circuit can be used to remove f_c thereby isolating one of the sidebands. Sideband modulation has better error performance in noisy environments at a cost of reduced usable bandwidth and relatively poor power efficiency. However, it can be implemented using inexpensive hardware. The comb like pattern of the carrier signal, load modulated with the baseband modulated subcarrier can be seen in the right hand side of Figure 3. It takes 128 carrier cycles (9.4 μs) to modulate a single bit.

Hancke notes that the distance resolution of a channel of bandwidth B is 'roughly' equal to c/B where c is the speed of light. According to this formula, the distance resolution is the propagation distance in one bit period. To clarify some of the roughness, we will examine this further: the distance resolution is actually a function of the timing resolution - the verifier stops the clock when the bit is received. But distance bounding protocols only need to exchange a single bit so the timing resolution is determined by whether the prover can *start* modulating the bit at an arbitrary point in time[4]. To make this distinction

[4] In the discussion that follows we assume that the prover's distance bounding logic has direct access to the physical layer to modulate its response as a single bit. This cannot be done within the existing ISO 14443 standard. Furthermore, ISO 14443 requires a card to begin modulating its response only after a defined frame delay time (FDT) has elapsed. This means that the first response bit has to be aligned to a bit grid defined by the FDT. In examining the principles of timing resolution, we ignore this restriction in ISO 14443.

more concrete, recall that the bit period of for ISO 14443 cards is 128 carrier cycles - which equates to a propagation distance of 2.8 km. However, because the prover is responding with a single bit sent on on a clear channel, she does not need to wait for some bandwidth determined bit boundary to be able to start modulating. At the instant the response is calculated the prover has the ability to *start* producing a modulation peak on the main 13.56 MHz carrier. Once chosen, the starting modulation cycle effectively determines the ending cycle which fixes when the verifier stops the timer. From the verifier's perspective, the response bit can start arriving on any cycle of the carrier. With some important qualifications which we will soon discuss, the timing resolution is closer to $1/f_c$ (74 ns or 22 m propagation). For ISO 14443 cards, this is a significantly finer resolution than the bit period alone would indicate. We will soon see that this can be improved even further if subcarrier modulation is not used.

Unfortunately, for more subtle reasons, this does not mean that usable timing resolution is purely a function of f_c and independent of the bit rate. To understand why this is so, we need to consider how a malicious adversary might manipulate a modulation scheme to their advantage to avoid detection of a relay attack. If a contactless smart card system is protected by a timing-based distance bounding protocol and an attacker wants to launch a relay attack against it in the style of Hancke [11] or Kfir [15], to be successful the attacker needs to absorb the delay that is introduced by the relay so that the round trip time falls in the range that the verifier will accept. If the bit period is sufficiently long, the attacker can begin modulating a guess of the response bit to the real verifier (reader) at the expected time via the rogue prover $\bar{A}$. If the delay introduced by the relay devices is not too great, then at some point part way through the modulation of the guessed bit the attacker will learn (via the rogue verifier $\bar{B}$ which relays the signals) the correct value by monitoring the response of the real card. If the guess is wrong, the part-modulated bit can be switched to the correct modulation pattern. If this switch occurs sufficiently early in the bit period, the inconsistency will be interpreted as noise and the new value will be accepted. This is possible because of the high degree of redundancy in the subcarrier modulation method. This redundancy also means that the attacker can read the card's response bit well before the end of the bit period by directly monitoring the amplitude modulation on f_c rather than accessing the data via processing a side band. To avoid this, we need to adopt a modulation scheme that is not susceptible to having a modulated bit changed part way through the bit period. The most effective way to do this is to reduce the number of carrier cycles that represent a bit period. This increases the channel bandwidth but also the susceptibility to bit errors through channel noise.

We have already noted that the timing resolution for a modulated channel is *potentially* $1/f_c$ and that for ISO 14443 cards at 13.56 MHz, this represents a timing resolution of 74 ns. The qualification is necessary because the resolution depends on how quickly the card can increase the load on the reader's antenna circuit to produce a detectable carrier amplitude change. This depends on the quality of the inductive coupling between the card and reader antenna loops.

The coupling quality degrades as the distance increases and it takes more cycles to produce a detectable modulation change, since the card can only apply a slowly increasing load. We have determined through experimental observation (see Section 5) that at closer distances, a card can produce a detectable modulation peak within a half cycle. This means that the timing resolution can be as low as $1/2f_c$ or 37 ns, the period between successive carrier peaks.

Rogue relay devices introduce detectable delays We have noted that the propagation delay for short range relay attacks is of the order of 10-20 ns. So at first glance it would appear that even in the best case (37 ns), a 13.56 MHz carrier will not be able to provide the required timing resolution. However, this overlooks a crucial observation - relay attacks introduce unavoidable delays beyond the time it takes the signal to propagate through the air. The rogue relay devices themselves each incur a circuit propagation delay known as *group delay* or *envelope delay* - the amount of time that the amplitude modulated signal is delayed by its passage through a device [14]. For example, Hancke's relay attack introduces a total delay of 15-20 μs for a relay distance of only 50 m. A signal could propagate over a 6 km round trip in this time. Given that the actual distance is 50 m, the delay contribution of the relay devices is clearly significant. Hancke's attack was a proof of concept that used inexpensive RF relay equipment with a high group delay of approximately 4 μs. Microwave transceivers operating at gigahertz frequencies can have group delay in the order of tens of nanoseconds [2]. This type of equipment is typically found in signals intelligence applications and is both expensive and exotic[5]. Less costly equipment will have a significantly higher group delay and will therefore be easier to detect.

The key reason why it may not be necessary to resort to UWB radio as Hancke has suggested [12] is due to the unavoidable group delay introduced by the relay devices. The relayed signal needs to pass through $\bar{A}$ and $\bar{B}$ in two directions so the group (circuit) delay of the relay devices introduces approximately 40 ns of cumulative delay. Because of this, we argue that a distance bounding protocol only needs to be able to detect an introduced delay of approximately 50 ns to still be able to detect highly sophisticated short range relay attacks by well funded attackers. With some limitations, this is achievable using amplitude modulation of a 13.56MHz carrier because positive or negative carrier peaks occur every 37ns. The introduced delay is detectable because it is greater than the period between carrier peaks.

In summary, the ISO 14443 card to reader modulation scheme is inappropriate for the time critical phase of a distance bounding protocol, not simply because of its low bit rate but because the large number of carrier cycles per bit makes it easy for an attacker to manipulate - by guessing the response bit and changing the modulation pattern part way through the bit period if the guess turns out to be wrong. Introduced relay delays for short range attacks (which

[5] The Macom SMR-4820 Compact Microwave Search Receiver claims a group delay of <15ns for 10MHz output bandwidth. See http://www.macom.com/sigint/PDF/4820.pdf for product specification sheet.

are the most difficult to detect) can be absorbed in this way. To address this shortcoming without resorting to the significant expense of implementing UWB radio in the smart card we propose the following modifications: to replace the subcarrier method of communication in the time critical phase with a simple peak detection approach on the main carrier. When the XOR response to the challenge is logical one, the card's load resistor is switched into circuit (via a special XOR instruction) and this produces an amplitude change in the reader antenna circuit. If the result is zero, the resistor is not switched and there is no change. Whilst being susceptible to noise errors, (which can be accommodated by adjusting the number of rounds), this scheme has very low latency and minimises the attacker's ability to guess and subsequently change the bit they are modulating. The card should implement the protocol in such a way that it calculates its response in the smallest possible fixed number of carrier cycles. The verifier will know how many cycles the card requires, allowing it to sample the amplitude of the carrier in a window starting at at the nominated cycle, to detect zero or one by the absence or presence of a peak. If the genuine card's responses are being communicated via a relay the introduced delay will push the modulation peak onto a later cycle. Note that side channel defense measures (active power consumption filters, random delay cycles etc.) would need to be disabled during the time critical phase. To protect the long term key, the card must introduce its own randomness into the bits that are XORed with the key.

5 Investigations into modulation latency

We hypothesized that a contactless smart card could alter its power consumption using existing load modulation circuitry to communicate in our proposed style of side channel leakage, with sufficiently low latency to detect sophisticated short range relay attacks by well funded attackers. To test the latency aspect of this hypothesis[6] we investigated the rate of change in carrier amplitude that a card could effect via load modulation, as detected in the receive circuit of the reader. As we have previously noted, the rate of change determines the effective timing resolution.

5.1 Experimental setup and rationale

We used a Philips Mifare Pegoda development kit reader which conforms to the ISO 14443 type A standard. A digital oscilloscope with a 200 MHz sampling rate was directly attached to the receive circuit of the reader's antenna. We captured multiple traces of the 'REQA' command/response sequence with three different models of card, all of type A (an advanced dual-interface microprocessor card, a 4K Mifare card and a 'Ultra Lite' low cost Mifare card). Figure 3 shows a complete sequence. REQA is the first command issued by the reader to the card

[6] Note that we have not implemented the proposed communication method or distance bounding protocol in an actual smart card. Such an implementation would require very low level changes to the card operating system mask and circuitry.

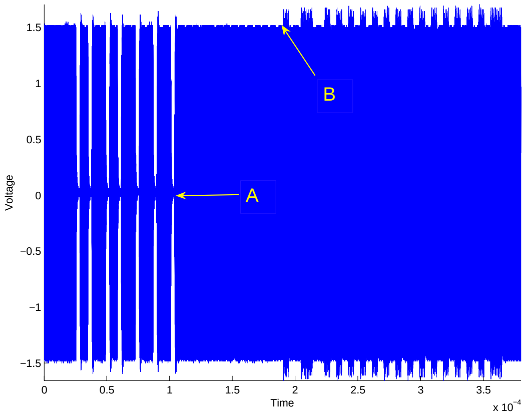

Fig. 3. REQA command/response sequence as measured in reader RX circuit.

when it comes into the reader's field. It was chosen because the standard requires the card to begin modulating its response a fixed number of carrier cycles after the reader's last command bit. This mirrors our proposal, where the card takes a fixed number of cycles to compute its response however, in the case of REQA, the response always begins with the modulation of a start bit - equivalent to a logical 'one' in our scheme.

Traces were recorded with cards at different distances and orientations to the reader. Orientation and distance effect the voltage across the reader's receive circuit. The larger the distance the greater the voltage and as the voltage increases, the rate of change that the card can effect on the carrier amplitude decreases. Hence greater card to reader distances mean more cycles are needed to produce a detectable change in carrier amplitude. This can be seen clearly in Figure 4 which presents the carrier for four different voltages around the time marked B in Figure 3. The first carrier half cycle that the card attempts to modulate is marked with a vertical line. For the two smallest voltage traces an amplitude change is clearly evident. For the highest voltage trace, there is no discernible amplitude change on this cycle - the first significant change can be seen one and a half cycles later at time 0.8729 μs. We found that the rate of amplitude change is a function of the voltage, irrespective of the orientation and distance that produced it, so in our analysis we consider dependent variables such as delay in detecting a modulation peak, against antenna voltage (rather than distance and orientation which in practice are difficult to specify with precision).

In ISO 14443 type A, the reader communicates with the card via 100% amplitude shift keying (short suspensions of the carrier) as can be seen in the left part of Figure 3. We aligned the traces on the first cycle that resumes the carrier in the last bit pause. This can be seen in Figure 5 which shows that the alignment is precise and unambiguous. Figure 5 presents a short period of the

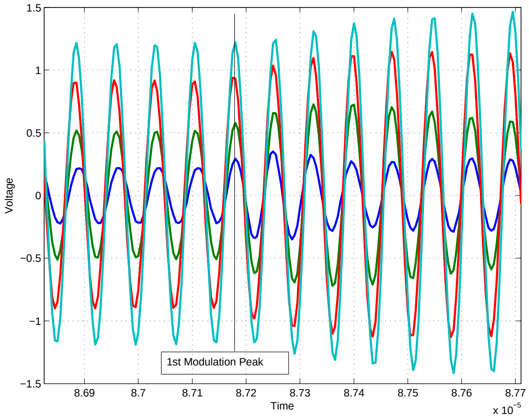

Fig. 4. Card communicates to reader via load modulation. Higher antenna voltages produce delays in effecting a detectable amplitude increase.

region marked at point A in Figure 3 for five different trace voltages. Using this alignment as a reference point, we can confidently identify the half cycle that the card starts its load modulation on irrespective of whether an amplitude change is actually evident. Since we can identify this starting half cycle, we can measure the number of cycles required to produce a detectable amplitude change and hence characterise the timing latency at that voltage.

5.2 Analysis and results

Using Matlab, (a numerical analysis and simulation tool) we developed a model of an amplitude peak detection circuit. The detector compares each half cycle peak amplitude to an average of eight previous peaks of the same sign. If the difference exceeds a threshold value, the current peak is signaled as a modulation. The choice of threshold value is important as it determines the sensitivity to detecting true modulation peaks and also the likelihood that channel noise will be falsely interpreted as a peak. The tradeoff between such true and false positives becomes more delicate as the antenna voltage increases because changes in amplitude become progressively less pronounced, finally disappearing into the noise floor. This effect can be seen in Figure 6. Each data series represents the percentage change in amplitude for a modulation peak number versus antenna voltage. For example, the first modulation peak shows a 30% amplitude increase when the antenna voltage is 200 mV but at 1200 mV there is barely any discernible increase. If the reader examines the amplitudes of each trace on the 'First Modulation Peak' line in Figure 4 in relation to the preceding peaks, it should be evident that the respective increases correspond to the Peak 1 series in Figure 6.

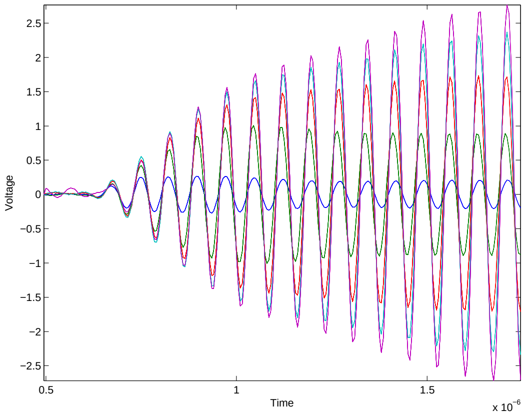

Fig. 5. Trace Alignment on resumption of carrier in last reader to card bit pause.

We derived a voltage-dependent function to generate the 'Threshold' values in Figure 6 in the following manner. Based on 1000 traces taken with one card type at a range of voltages, we experimentally identified a threshold for each trace (by choosing successively smaller values) that correctly identified the modulation peaks whilst keeping false positives below a maximum of 20. Lowering the threshold increases the probability of detecting the real modulation cycles as early as possible but also increases the number of false detections of non-modulated peaks that have slightly higher amplitudes due to environmental noise. The peak detector examines approximately 2300 peaks per trace (the portion of the carrier signal between the points marked A and B in Figure 3) so the somewhat arbitrary false positive count of <20 loosely approximates an false detect error rate of $< 1\%$, based on the following rationale: in our proposal, the card signals zero by not modulating, which is what the card is doing in the 1150 cycles between points A and B so a threshold that produces no more than 20 false positives in this region provides an approximation of a false detect error rate on any individual peak of $< 1\%$ when the card sends zero (by not modulating).

The experimentally identified threshold/voltage data pairs were fitted to a third degree polynomial and this function was used to specify threshold values for traces at a range of voltages for the other two card types. We found that the voltage dependent threshold function derived from the data for one card produced very similar true and false positive rates when used on traces for the other cards. While we do not claim our results are conclusive in this respect due to the small number of cards that we have examined, it appears possible that a single voltage dependent threshold function will work across a range of cards. This would certainly simplify the implementation of the peak detection circuitry

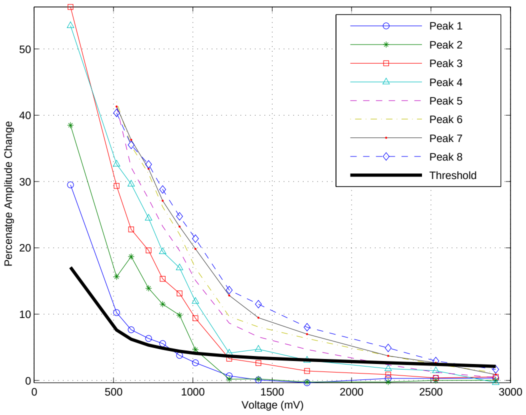

Fig. 6. Percentage change in modulation peak amplitude vs. antenna voltage.

in reader devices as a simple lookup table could supply the detection threshold for a given voltage.

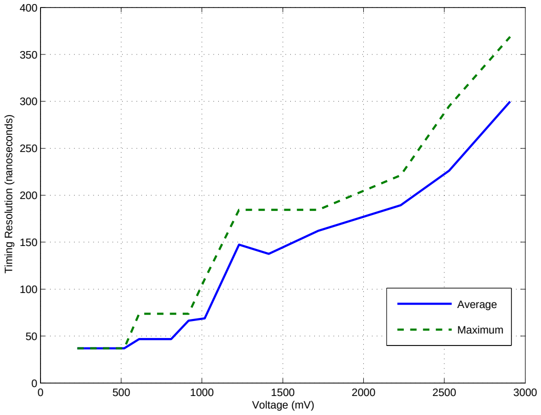

Fig. 7. Timing resolution using proposed method.

Figure 7 shows the average and maximum timing resolution as a function of antenna voltage. The first modulation peak is reliably detectable up to 500mV, so there is no detection delay, thus the timing resolution is $1/2f_c$ or 37 ns. Between 500 and 800 mV there is a slight increase in the average because the first peak is detected in only 75% of cases. We argue that the resolution up to 800mV is sufficient for the worst case scenario - detection of short range attacks using exotic and expensive relay equipment with very low group delay. The average timing resolution at higher voltages increases to 300 ns which is still 50 times smaller than the delay introduced by Hancke's proof of concept attack [11].

Attaining sub-50 ns resolution places a significant restriction on the card to reader operating distance. With our reader, the card needs to be within a few millimeters to operate in this voltage range. In practice, this would mean that the user would need to touch the card on the reader. This reduction in operating range is clearly a disadvantage though it is worth noting that such fine resolution is only required to detect short range attacks using sophisticated and expensive relay equipment. Presumably, a large payoff would be required to motivate an attacker to go to such effort and expense, perhaps larger value contactless payment applications or high security physical access control. In these higher risk application scenarios it may be quite sensible to make the act of using the card more overt and deliberate by requiring very close proximity operation. With our reader, 300 ns resolution was attained in the 4-5 cm range. Better distance performance may be possible and further investigation is required to assess usable operating range. We suspect that our reader may have been down on power as it was powered from a USB port. It could not operate a card at 10 cm, the maximum distance required by the standard. 7.5 cm was the limit of reliable operation.

6 Future work

Further investigation is needed into the impact of different RF noise environments on modulation and detection performance. While our experiments were carried out in a busy electrical engineering laboratory, this does not characterise the broad range of possible deployment scenarios.

Data needs to be gathered for a much larger range of readers and cards. Though we have not investigated it, we suspect that ISO 14443 type B cards may have better modulation performance at larger distances. Type B does not use 100% ASK in reader to card communications as type A does. Type A cards need a store of energy (capacitance) to continue operating during the short periods of carrier suspension. The availability of this reserve of power reduces the load that the card can apply to the antenna circuit to modulate the carrier. Interestingly, the smart card manufacturer Infineon has recently applied for a patent [13] on a decoupling circuit that makes the power reservoir selectively unavailable when the card is modulating. They claim that the circuit improves the modulation performance at larger distances. It would be worth investigating whether this innovation can further improve the operating range and effective timing resolution for type A cards using our proposed method.

7 Conclusion

We have proposed the first symmetric key based distance-bounding protocol that is resistant to so called ‘terrorist fraud’, together with a security analysis. In contrast to previous proposals the protocol is appropriate for implementation in resource constrained devices due to its computationally efficiency.

We have analysed the unconventional requirements that distance bounding protocols place on the communication channel used in the time critical phase, highlighting the importance of low latency as opposed to raw bit rate. In response to these requirements, we proposed a novel approach to communication that leverages the phenomena of side channel leakage, heretofore considered exclusively as a security vulnerability. We exploit the extremely low latency of side channel leakage to address the requirements of distance bounding protocols.

We presented experimental results indicating that a modified form of load modulation, used in the style of our proposed side channel leakage communication technique, can provide sufficient distance resolution to detect advanced relay attacks on ISO 14443 smart cards. Our technique has the disadvantage of reducing the operating range of the smart card. We have argued on the basis of these results, that it may not be necessary to incur the additional expense and complexity of implementing ultra wide band radio for the distance bounding communication channel, as Hancke and Kuhn [12] have asserted.

References

1. ISO/IEC 14443 Identification cards-contactless integrated circuit(s) cards-proximity cards. International Organisation for Standardisation, Geneva.
2. David Ballo. Measuring absolute group delay of multistage converters. In *Proceedings of 33rd European Microwave Conference*, volume 1, pages 89–92. IEEE, 2003.
3. Mihir Bellare, Ran Canetti, and Hugo Krawczyk. HMAC: Keyed-hashing for message authentication. Internet Request for Comment RFC 2104, Internet Engineering Task Force, February 1997.
4. Stefan Brands and David Chaum. Distance-bounding protocols. In *EUROCRYPT '93: Workshop on the theory and application of cryptographic techniques on Advances in cryptology*, pages 344–359, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc.
5. L. Bussard. *Trust Establishment Protocols for Communicating Devices*. PhD thesis, Institut Eurécom, Télécom, Paris, 2004.
6. Srdjan Capkun, Levente Buttyan, and Jean-Pierre Hubaux. Sector: secure tracking of node encounters in multi-hop wireless networks. In *SASN '03: Proceedings of the 1st ACM workshop on Security of ad hoc and sensor networks*, pages 21–32, New York, NY, USA, 2003. ACM Press.
7. Srdjan Capkun and Jean-Pierre Hubaux. Secure positioning in wireless networks. *IEEE Journal on Selected Areas in Communications: Special Issue on Security in Wireless Ad Hoc Networks*, 2006. To appear.
8. Yvo Desmedt. Major security problems with the 'unforgeable' (Feige)-Fiat-Shamir proofs of identity and how to overcome them. In *SecuriCom '88*, pages 15–17, SEDEP Paris, France, 1988.

9. Klaus Finkenzeller. *RFID Handbook*. John Wiley and Sons, Hoboken, NJ, 2nd edition, 2003.
10. Lishoy Francis, William G. Sirett, Keith Mayes, and Konstantinos Markantonakis. Countermeasures for attacks on satellite tv cards using open receivers. In *CRPIT '44: Proceedings of the 2005 Australasian workshop on Grid computing and e-research*, pages 153–158, 2005.
11. Gerhard Hancke. A practical relay attack on ISO 14443 proximity cards. Manuscript, February 2005. Available at http://www.cl.cam.ac.uk/g̃h275/relay.pdf accessed October 2005.
12. Gerhard Hancke and Markus Kuhn. An RFID distance bounding protocol. In *Proceedings of the IEEE, SecureComm 2005*, September 2005.
13. Infineon. Device and method for supplying a data transfer unit with energy. US Patent Application 20050252972, 17 November 2005.
14. Adrian Jones and Jason McManus. The measurement of group delay using a microwave system analyser. *Microwave Journal*, 43(8):106–113, 2000.
15. Ziv Kfir and Avishai Wool. Picking virtual pockets using relay attacks on contactless smartcard systems. In *Conference on Security and Privacy for Emerging Areas in Communication Networks – SecureComm 2005*, Athens, Greece, September 2005. IEEE. To appear. Available at http://eprint.iacr.org/2005/052.pdf accessed September 2005.
16. Paul Kocher, Joshua Jaffe, and Benjamin Jun. Differential power analysis. In Michael Wiener, editor, *Advances in Cryptology – CRYPTO '99*, number 1666 in Lecture Notes in Computer Science, pages 399–397. International Association for Cryptologic Research, Springer-Verlag, Berlin Germany, August 1999.
17. Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone. *Handbook of applied cryptography*. CRC Press series on discrete mathematics and its applications. CRC Press, 1997. ISBN 0-8493-8523-7.
18. Naveen Sastry, Umesh Shankar, and David Wagner. Secure verification of location claims. In *WiSe '03: Proceedings of the 2003 ACM workshop on Wireless security*, pages 1–10, New York, NY, USA, 2003. ACM Press.