Kerckhoffs in Prison: A Study of the Steganalyst's Knowledge

Martin Beneš
Universität Innsbruck
martin.benes@uibk.ac.at

Rainer Böhme
Universität Innsbruck
rainer.boehme@uibk.ac.at

***Abstract*—The goal of a steganalyst is to distinguish stego objects sent by a steganographer with a secret message hidden inside from cover objects with no meaningful message. The steganalyst is only as good as her knowledge of the sender's choices during the embedding. To create a security margin for the real world, security definitions often borrow Kerckhoffs' principle, i.e., they assume that the attacker knows the system. However, a strict interpretation of the principle, "only the key is secret," does not apply to steganography because a public cover would compromise security. In this paper, we study two interpretations of Kerckhoffs' principle for steganography using a conceptual framework of the steganalyst's knowledge. The framework connects both interpretations and encompasses other existing notions of knowledge in the literature. Viewing access to the cover as a side-channel attack allows us to adapt Kerckhoffs' principle to Simmons' prisoners' problem.**

***Index Terms*—steganalysis, Kerckhoffs' principle, prisoners' problem, cover-source mismatch, side channel**

I. Introduction

The prisoners' problem refers to the scenario of two prisoners who want to communicate secretly, without raising suspicion of a warden called the attacker. The solution to the problem is steganography: the sender hides the secret message in an inconspicuous cover object using a secret key, shared with the recipient [1]. The attacker, who does not know the key, tries to distinguish between an innocent cover and a stego object containing a secret message.

The indistinguishability of cover and stego, a common metric of steganographic security [2], is strongly impacted by the knowledge the attacker has of the sender. According to Kerckhoffs' principle [3], "*it should not cause trouble were it [the system] to fall into enemy hands*" (translation from [4, p. 4]). Its common interpretation is that the method should be secure even if everything but the key is public [5], [6]. In steganography it means the following:

Kerckhoffs' principle (KP) The security of a stego-system is based solely on the secret key.

The KP attacker is too strong for steganography, because access to the cover allows for near-perfect detection by comparing it to the object under analysis. The common patch is to require that the cover instance is secret, but grant the attacker access to the sender's cover source [7], [8], [9, Ch. 1]. This corresponds to:

Kerckhoffs-in-prison principle (KPP) The security of the stego-system is based on the shared secret key and a secret cover object sampled by the sender.

KPP reduces the security margin, which may even be negative, e.g., if the attacker has information about the cover instance. Making such compromises goes against the original idea of Kerckhoffs' principle, and leaves its adaptation to the prisoners' problem unsolved [10], [11].

This study proposes a solution to the dilemma of secret covers under Kerckhoffs' principle by relating it to side-channel attacks known from cryptography [12]. It unifies existing notions of the attacker's knowledge by proposing a conceptual model of the sender's choices. This sender model can express assumptions about the attacker formally and graphically, which we demonstrate on selected literature.

This paper is divided as follows: Section II summarizes related work. Section III introduces the sender model. Section IV applies it to express the attacker's knowledge, describes the common types of attackers, and visualizes their knowledge graphically. Section V shows how the sender model relates to the existing notions of attacker's knowledge in steganalysis. The final Section VI provides a discussion.

Conventions: In this paper, realizations are written in lowercase, x, random variables in uppercase, X, and sets in calligraphic font, $\mathcal{X}$. Random variables are defined over finite sets, whose elements can be scalars, vectors, or functions. Let $P(X = x)$ be the value of the distribution function of the random variable X at point x. The cover object is denoted $x^{(0)}$ and the stego object $x^{(m)}$ with a message m of size $|m|$. An unknown object (either cover or stego) is denoted $x^{(y)}$, $y \in \{0, m\}$. Mock values chosen by the attacker are marked with prime, e.g., m'.

II. Related work

Attacker's knowledge: Steganography has dominated secret communication for most of history, with the well-known examples of scalp tattoos or wax tables [13]. Such methods are secure as long as the attacker does not know where to look. The idea to abandon security by obscurity was first postulated by Kerckhoffs' in 1883 [3]. Claude Shannon phrased Kerkchoffs' principle as *only the key is secret* [5].

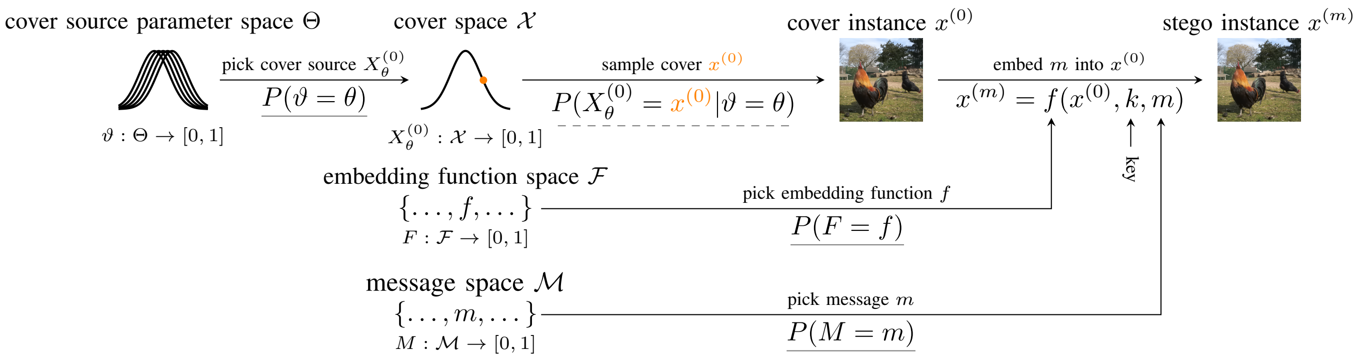
cover source parameter space Θ
pick cover source $X_\theta^{(0)}$
$P(\vartheta=\theta)$
$\vartheta:\Theta\to[0,1]$
cover space $\mathcal{X}$
sample cover $x^{(0)}$
$P(X_\theta^{(0)}=x^{(0)}|\vartheta=\theta)$
$X_\theta^{(0)}:\mathcal{X}\to[0,1]$
cover instance $x^{(0)}$
embed m into $x^{(0)}$
$x^{(m)}=f(x^{(0)},k,m)$
key
stego instance $x^{(m)}$
embedding function space $\mathcal{F}$
$\{\dots,f,\dots\}$
$F:\mathcal{F}\to[0,1]$
pick embedding function f
$P(F=f)$
message space $\mathcal{M}$
$\{\dots,m,\dots\}$
$M:\mathcal{M}\to[0,1]$
pick message m
$P(M=m)$

Fig. 1. Sender model. Underline means known under KPP und KP, dashed underline means known under KP only. The spaces Θ, $\mathcal{X}$, $\mathcal{F}$, and $\mathcal{M}$ are public.

In 1976, Diffie and Hellman established a systematization of attacks based on what an attacker possesses or knows [14].

The work on the attacker's knowledge in steganalysis and the related field of watermarking mostly builds on the Diffie–Hellman terminology [15], [16], [17], [10]. Cachin's seminal information-theoretic model of steganographic security [2] was extended by Katzenbeisser [18] to encompass computational aspects using security games with steganalysis-specific oracles. Kerckhoffs' principle has been addressed by Ker, who lists four interpretations for batch steganography [4]. Böhme studies steganography epistemologically and proposes conditional cover models to incorporate attacker knowledge [11].

Side-channel attacks: Side-channel attacks refer to the attacker gaining knowledge of the communicating parties' execution environment, e.g., timing [12], cache accesses [19], or power consumption [20], which substantially increases the attacker's advantage. In cryptography, side-channels attacks recently led to the realization that Kerckhoffs' principle does not represent the strongest attacker possible [21, p. 8], [6], [22]. In steganography, side-channel attacks w.r.t. to the execution environment of the sender have not yet been studied.

III. Sender model

This section lays the foundation for exploring the attacker's knowledge. It defines a model of the sender's choices, which is illustrated in Fig. 1 and explained in the following.

a) Setup phase: The steganographic communication is preceded by a setup phase in which the sender and recipient agree on a key k and an embedding function f. The function f is drawn according to the distribution $P(F=f)$ from a space $\mathcal{F}$, e. g., the available steganographic applications.

b) Embedding: We model the cover selection as a dual-stochastic process. The sender chooses the cover source $X_\theta^{(0)}$ according to $P(\vartheta=\theta)$, e.g., by choosing a capturing device. The parameter θ completely describes the cover source: the processing steps, their order, and the parameters' values. Then, a cover instance $x^{(0)}$ is drawn from the cover source $X_\theta^{(0)}$ according to $P(X_\theta^{(0)}=x^{(0)}|\vartheta=\theta)$, e.g., by taking a photo. In general, the cover is a sequence of $|x^{(0)}|$ elements.

The message is determined by the sender, but can be thought of as being drawn from a space $\mathcal{M}$ according to $P(M=m)$. We assume F, M and $X^{(0)}$ to be independent, which may not always hold in practice [23]. For brevity, $P(F=f)$ and $P(M=m)$ are written jointly as $P(F=f,M=m)$.

Finally, the sender embeds the message m in the cover $x^{(0)}$ using the embedding function f and the key k and sends the resulting stego $x^{(m)}$ to the recipient over a channel observed by the attacker.

c) Extraction: The recipient receives $x^{(m)}$ and extracts the message $m=\bar{f}(x^{(m)},k)$. $\bar{f}$ is the extraction function corresponding to the embedding function f. The recipient extracts "blindly", i.e., without knowing the cover instance.

IV. Attacker's knowledge

We turn our attention to the attacker. Section IV-A shows how the sender model serves to express the attacker's knowledge. Section IV-B lists the common attackers. Section IV-C defines the steganographic side-channel attacks, which connect the two interpretations of Kerckhoffs' principle. Finally, Sec. IV-D visualizes the attacker's knowledge graphically.

A. Attacker's knowledge via sender model

The strength of the attacker grows with her knowledge of the sender's choices of the cover source $X_\theta^{(0)}$, the cover instance $x^{(0)}$, the embedding function f, the message m, and the key k. The attacker's knowledge can be modeled as the joint probability $P(\Omega=\omega)$, where $\Omega=(\vartheta,X_\theta^{(0)},F,M,K)$ and $\omega=(\theta,x^{(0)},f,m,k)$. The independence assumptions ($\perp$) from the sender model in Sec. III allow us to decompose this probability into

$$\begin{aligned}P(\Omega=\omega)\overset{\perp}{=}&\,P(K=k)P(F=f,M=m)\cdot\\&P(X_\theta^{(0)}=x^{(0)}|\vartheta=\theta)P(\vartheta=\theta).\end{aligned}\tag{1}$$

The model introduces formal notation for attackers of different strengths. Observing an object on the channel $x^{(y)}$ increases the attacker's knowledge, $P(\Omega=\omega|x^{(y)})\geq P(\Omega=\omega)$, because $x^{(y)}$ carries information about the sender's choices.

B. Types of attackers by knowledge

Now we elaborate on the most common types of attackers, in order of increasing knowledge. We use expressions in square brackets to indicate the attackers' knowledge as follows:

the underlined symbols are exactly known, while the non-underlined are not. In fact, the attacker may still have prior knowledge about the unknown symbols, but we assume that it is not feasible to use it for steganalysis, e.g., because it might require exhausting too large a search space.

a) Realistic attacker $[X_\theta^{(0)}\ m\ f\ x^{(0)}\ k]$*:* In steganalysis, a "realistic" attacker is a synonym for a weak attacker [24]. Such an attacker is sub-KPP and, depending on the context, does not know the cover source $X_\theta^{(0)}$, the embedding function f, the message m, or a combination of these.

b) Cover-source oracle $[\underline{X_\theta^{(0)}}\ m\ f\ x^{(0)}\ k]$*:* An attacker with a cover-source oracle (CSO) can draw a number of covers from the same cover source as the sender, $x^{(0)'} \sim X_\theta^{(0)}$. This number may be fixed, e.g., limited by the organizers of a competition who provide training data. Lack of access to a cover-source oracle causes cover–source mismatch [25].

c) White-box embedding $[X_\theta^{(0)}\ m\ \underline{f}\ x^{(0)}\ k]$*:* An attacker with white-box embedding (WBE) can generate the stego object $x^{(m')'}$ for a cover $x^{(0)'} \sim X_{\theta'}^{(0)}$, message m', and a key k' of her choice. Such an attacker does not necessarily know the cover source $X_\theta^{(0)}$ or the message m of the sender.

We do not call this "oracle" because the sender's key is not used for embedding. This distinguishes our notion from the oracle defined in [18]. Lack of access to WBE is also known as stego–scheme mismatch [25].

About the message size: The message size is a special piece of knowledge in steganalysis. Compatibility attacks [26] demonstrate that in principle a single bit change could lead to perfect detection. However, this requires knowledge of local discontinuities in the high-dimensional source distribution, which is often not available. Many detectors, in particular learning-based ones, approximate the volume of covers by a smooth manifold. As longer messages tend to require more changes to the cover, such stego objects are more distant from covers and thus better separable. This is why the message size is relevant in practice. Learning-based steganalysis often trains for a fixed $|m|$ [27], [28] and is evaluated under the assumption that the attacker knows $|m|$ exactly. How to relax this assumption has been explored in [29].

d) KPP attacker $[\underline{X_\theta^{(0)}\ m\ f}\ x^{(0)}\ k]$*:* The Kerckhoffs-in-prison principle combines CSO, WBE, and a known message m [30, Ch. 3]. The knowledge of the KPP attacker is

$$P(\Omega = \omega) \overset{KPP}{=} P(X_\theta^{(0)} = x^{(0)} | \vartheta = \theta) P(K = k). \qquad (2)$$

Research often assumes a *quasi-KPP* attacker with CSO, WBE, and known message size $|m|$. The impact of known m on steganographic security is yet to be studied, but it is likely small, as the message is usually permuted using a pseudo-random sequence determined by the key k.

e) KP attacker $[\underline{X_\theta^{(0)}\ m\ f\ x^{(0)}}\ k]$*:* Kerckhoffs' principle represents the KPP attacker who knows the cover $x^{(0)}$. The knowledge of the KP attacker is

$$P(\Omega = \omega) \overset{KP}{=} P(K = k). \qquad (3)$$

The KP attacker detects steganography near-perfectly: for LSB steganography, the detection error is $2^{-|m|}$. This is the strongest attack model considered here.[1]

C. Steganographic side-channel attacks

Steganography and watermarking research has struggled with the interpretation of Kerckhoffs' principle due to the existence of a cover or host signal, respectively. With blind extraction, the cover instance can be seen as an internal state of the sender. Thus, access to the cover is a steganographic side channel. While side-channel attacks in cryptography imply attackers stronger than assumed by Kerckhoffs' principle, a perfect cover side channel in steganography corresponds to the notion of a KP attacker.

In reality, covers are not an internal state [8]. Thus, attackers can be stronger than KPP by acquiring partial information about the cover, such as the cover thumbnail [34]. Some information about the cover is also carried in the stego object, when the elements are dependent, e.g., image pixels. It has been shown many times that the cover can be estimated from the stego using calibration [35], [36], [37] or cover predictors [38], [39]. In general, the KP attacker is the upper bound for the steganographic side-channel attacks with respect to the cover.

D. Connecting the knowledge

Figure 2 visually combines everything said so far. In the center of the diagram is the KPP attacker. To the right, she is connected with the KP attacker by side-channel attacks, related to learning more about the cover instance $x^{(0)}$. To the left is a coordinate system, representing the knowledge of the cover source, $P(\vartheta = \theta)$, and the knowledge of the message and the embedding function, $P(M = m, F = f)$. The dotted lines are variants of white-box embedding, vanilla (WBE), with known message size (WBE+$|m|$), with known message (WBE+m), as well as the cover-source oracle (CSO). The KPP attacker is located in $(1, 1)$, and the quasi-KPP attacker is at the intersection of WBE+$|m|$ with CSO. In the origin of the coordinate system is the (theoretical) attacker without any knowledge of the cover source and embedding.

We have selected examples from the literature which represent attackers with different level of knowledge and mapped them onto Fig. 2. This provides context and allows comparing assumptions between the papers. Note that some setups may be difficult to be mapped onto the plane, e.g., sub-KPP attackers with partial information about cover.

V. Examples in the literature

Finally, we show that the sender model can capture existing notions of the attacker's knowledge from the literature. In contrast to Sec. II, which summarizes the related work, here we draw a direct comparison with the sender model.

[1] In theory, it is possible to construct even stronger attackers, e. g., when the key is too short [31] or being reused [32], [33].

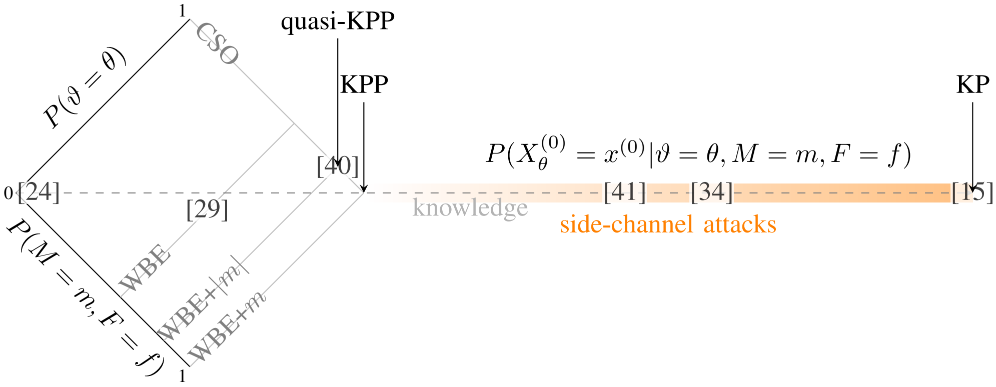

Fig. 2. Mapping assumptions in the literature to a graphical representation of the attacker's knowledge. The Kerckhoffs-in-prison principle (KPP) attacker has the cover-source oracle, the white-box embedding, and the message. The Kerckhoffs' principle (KP) attacker knows the cover instance $x^{(0)}$. BOSS [40] assumes the white-box embedding of HUGO. [29] works with WBE without knowledge of $|m|$. Heterogeneity during the ALASKA challenge [24] made the attacker weaker. Steganalysis with cover thumbnail [34] or with LRT [41] assume super-KPP attackers with partial knowledge of the cover instance. One of the attacks in [15] assumes knowledge of the pre-cover instance.

A. *Cachin's steganographic security*

A fundamental concept in steganalysis is Cachin's steganographic security [2]. Security is expressed in terms of ability to distinguish whether an input $x^{(y)}$ was drawn from the cover source, $x^{(0)} \sim P_C$, or from the stego source $x^{(m)} \sim P_S$. Cachin assumes the KPP attacker: "*The probability distributions are assumed to be known to all parties if not stated otherwise*" [2, p. 5]. The distinguishability is expressed via the likelihood-ratio test with a threshold T,

$$\Lambda(x) = \frac{P_C}{P_S} = \frac{P(X_\theta^{(0)} = x^{(y)}|\vartheta = \theta)}{P(X_\theta^{(m)} = x^{(y)}|\vartheta = \theta, F = f, \forall m \in M)} \geq T. \tag{4}$$

(We slightly abuse the notation in the denominator to allow for any combination of message and key.)

The stego source P_S is a relative term. For the sender who knows the cover $x^{(0)}$ and the key k, the stego object $x^{(m)}$ is a deterministic projection of the cover. From the perspective of the attacker, who lacks this knowledge, the stego appears to be non-deterministic, drawn from a source.

Cachin's security inspired model-based steganography [42], [27]. In MiPOD, the cover and stego sources are modelled by Gaussian distributions. The MiPOD-like likelihood-ratio test (LRT) is often used to benchmark steganography [43], [41]. Steganalysis with the MiPOD LRT assumes a super-KPP attacker with access to the cover variance.

B. *Diffie–Hellman categorization*

Table I shows how to adapt the Diffie-Hellman attacks (DHA) [14] to steganalysis, similarly to watermarking [17]. DHA assume that the stego objects reuse the key, as in [18].

In some schemes, reusing the key causes a constant fingerprint that can be estimated with enough captured stego objects [44], [32]. Under these circumstances, the attacker can circumvent the sender model and attack $P(K = k)$ directly. The solution to increase security is to combine the key with an initialization vector (IV) derived from the cover [45]. The method to derive the IV must be invariant w.r.t. f if the recipient extracts blindly and is known to the WBE attacker.

TABLE I
ADAPTATION OF DIFFIE–HELLMAN ATTACKS TO STEGANOGRAPHY.

Attack	Attacker has
Ciphertext-only (CO)	stego objects
Known-plaintext attack (KPA)	cover–stego pairs
Chosen-plaintext attack (CPA)	stego objects to chosen covers

In general, DHA adapted for steganography lack expressive power. For example, they cannot capture sub-KPP attackers, e.g., under cover-source mismatch.

C. *Franz and Pfitzmann's framework*

Inspired by the DHA, Franz and Pfitzmann categorize attacks [15], as shown in Tab. II, and propose natural steganography as a remedy. Their attacks can capture perfect knowledge of the pre-cover or the message, which goes beyond the DHA. Furthermore, the work discusses passive and active attacks, which our sender model omits. The differences between capturing devices are explored, but the attacks stay within the KPP: "*We have to assume that the attacker knows the scan process*" [15, p. 7]. Generally, this work models knowledge as a binary state and does not allow for partial information.

TABLE II
EXPRESSING FRANZ' ATTACKS VIA THE SENDER MODEL.

Attack	Sender model	Strength
stego-only-attack	full $P(\Omega = \omega)$	CSO
emb-stego-attack	$P(M = m) = 1$	KPP
cover-stego-attack	$P(X_\theta^{(0)} = x^{(0)}\|\vartheta = \theta) = 1$	side channel
cover-stego-emb-attack	$P(M = m) = 1$ $P(X_\theta^{(0)} = x^{(0)}\|\vartheta = \theta) = 1$	KP

TABLE III
EXPRESSING KER'S INTERPRETATIONS OF KERCKHOFFS' PRINCIPLE FOR BATCH STEGANOGRAPHY VIA THE SENDER MODEL.

	Description	Sender model
a)	Does not know the total message size	$P(\lvert m\rvert) = \frac{1}{\lvert\mathcal{M}\rvert}$
b)	Knows the total message size, but nothing about the strategy	$P(\lvert m\rvert) = 1$ $P(F = f) = \frac{1}{\lvert\mathcal{F}\rvert}$
c)	Knows the per-sample message sizes but not which sample contains what	$P(\{\lvert m\rvert\}_{1:B}) = \frac{1}{B!}$
d)	Knows the strategy	$P(\{\lvert m\rvert\}_{1:B}) = 1$ $P(F = f) = 1$

D. Pooled steganalysis

Batch steganography extends the conventional scenario of one object under analysis to a batch of B objects, $\{x^{(y)}\}_{1:B}$ [44]. This involves spreading variable-length message chunks $\{m\}_{1:B}$ between the objects according to a spreading strategy, which is a part of f [46]. The scenario may also assume multiple senders, some of whom use steganography [23]. Batch steganography opens a wide combination of possible options for the attacker's knowledge. The adequate response of the attacker is pooled steganalysis, i.e., to collect evidence from all the objects in batch into a single decision.

Ker proposes several possible interpretations of Kerckhoffs' principle for batch steganography [4], shown in Tab. III. The interpretations suggest that knowledge of the message size suffices to uphold Kerckhoffs' principle. The interpretation d) is said to be "*probably too strong*", however, it is in fact d) that is the quasi-KPP.

An attempt to integrate batch steganography into our framework could add an object axis to the sender model, as sketched in Figure 3. Such an extension could capture the knowledge about each sender and each object in a batch, or combine their information. However, we are not in a position to define how the canonical attacker models should be extended besides the trivial cases, where every instance is KP or KPP.

E. Generative steganography

An alternative to steganography by cover modification (SCM) is to embed the message while synthesizing and object [47, Ch. 4.2]. This so-called generative steganography (GS) is a promising direction for the future of the field, especially natural-language steganography [48], [49]. Note that the steganalysts in SCM and GS solve very different tasks: the SCM attacker decides whether a natural object was modified, while the GS attacker has to distinguish whether a synthesized object was generated from one or another source of pseudo-randomness. The attackers against GS can possess knowledge about the embedding function f (incl. weights and parameters), the message m, and the key k. As there is no cover, the KPP attacker and the KP attacker are identical.

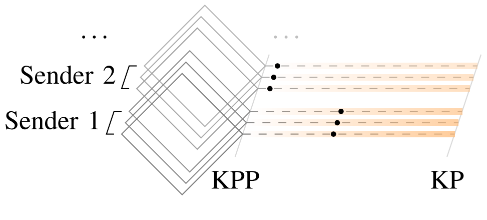

Fig. 3. Adapted sender model to batch steganography with multiple senders.

F. Selection-channel awareness

Selection-channel awareness (SCA) is a technique to pass the probability map[2] along with the stego object into a learning-based model in order to improve its performance [51], [52]. To obtain the probability map, the attacker must know the message size $|m|$ and the cover embedding costs, an intermediate value of the embedding function f. In practice, the embedding costs of the input $x^{(y)}$ are good estimates of the cover embedding costs. Therefore, SCA requires that the attacker has WBE with a known message size [9, Ch. 2].

G. Cover-source and stego-scheme mismatch

Cover-source mismatch (CSM) is caused by imperfect information about the cover source of the analyzed object, which has a negative impact on the performance of learning-based detectors. Attackers without the cover-source oracle may be affected by CSM, if they train on a wrong source. Uncertainty about θ can be mitigated to some extent, e.g., using atomistic or holistic strategies. Attackers with the cover-source oracle know θ, and are not affected by the CSM. A similar problem, called stego-scheme mismatch (SSM), occurs if f or $|m|$ is unknown [25].

The example of CSM serves as a counter-argument against blindly demanding Kerckhoffs' principle in experimental design. Despite being sub-KPP, studying CSM is highly relevant for the practical application of steganalysis, as demonstrated during the BOSS and ALASKA competitions [40], [24].

VI. DISCUSSION

Summary: The proposed sender model can be useful for a fair evaluation of steganographic security, because it allows us to compare the knowledge of different attackers. While rooted in the literature, it extends previous approaches in several ways. First, we believe it is the most comprehensive. Second, it also applies to learning-based detectors, where the attacker's knowledge is set indirectly through the composition of the training data and the choice of the model architecture. Third, it includes the relation to side-channel attacks, which are considered to be post-Kerckhoffs' in cryptography.

Limitations: This paper focuses on the knowledge of the attacker. It is silent about how this knowledge can be turned into a detection advantage. Our visualization of the sender model simplifies complex constructions such as cover sources or embedding functions. Any such projection omits points in the true space of attacker models. For example, we describe side-channel attacks as points between KPP and KP, whereas in practice it could happen that sub-KPP attackers

[2] [50] defines *selection channel* as the random path over the sets of elements, i.e., rather connected with $P(K = k)$.

learn partial information of the cover instance. Our formalisms make compromises for brevity. A full Bayesian treatment of knowledge and updates is left for future work.

Future work: The relationship between attacker knowledge and detection success, and hence steganographic security, could be studied both theoretically and experimentally for popular detectors, in particular those based on learning. Future work could also extend the sender model to active attackers, or identify canonical attacker models for pooled steganalysis. Yet another direction is to study how the sender's knowledge of the attacker's knowledge affects her choice. This becomes a game-theoretic problem, related to what has been studied for knowledge about embedding positions [53].

Acknowledgment

We thank Patrick Bas, Luca Campa, Benedikt Lorch, Nora Hofer, and the anonymous reviewers for their helpful comments. This work is partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021687.

References

[1] G. Simmons, "The prisoners' problem and the subliminal channel," in *Advances in Cryptology*. Springer, 1984, pp. 51–67.

[2] C. Cachin, "An information-theoretic model for steganography," in *IH*, vol. 1525. Springer, 1998, pp. 306–318.

[3] A. Kerckhoffs, "La cryptographie militaire," *Journal des Sciences Militaires*, vol. 9, pp. 5–38, 1883.

[4] A. Ker, "Perturbation hiding and the batch steganography problem," in *IH*. Springer, 2008, pp. 45–59.

[5] C. Shannon, "Communication theory of secrecy systems," *The Bell system technical journal*, vol. 28, no. 4, pp. 656–715, 1949.

[6] T. Knoll, "Adapting Kerckhoffs's principle," *Advanced Microkernel Operating Systems*, vol. 93, 2018.

[7] T. Filler, A. Ker, and J. Fridrich, "The square root law of steganographic capacity for markov covers," in *MFS*, vol. 7254. SPIE, 2009, pp. 62–72.

[8] A. Ker, "The square root law in stegosystems with imperfect information," in *IH*. Springer, 2010, pp. 145–160.

[9] T. Denemark, "Side-information for steganography design and detection," Ph.D. dissertation, Binghamton University, 2018.

[10] F. Cayre and P. Bas, "Kerckhoffs-based embedding security classes for WOA data hiding," *IEEE TIFS*, vol. 3, no. 1, pp. 1–15, 2008.

[11] R. Böhme, "An epistemological approach to steganography," in *IH*. Springer, 2009, pp. 15–30.

[12] P. Kocher, "Timing attacks on implementations of diffie-hellman, rsa, dss, and other systems," in *CRYPTO*. Springer, 1996, pp. 104–113.

[13] D. Kahn, "The history of steganography," in *IH*. Springer, 1996, pp. 1–5.

[14] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE TIT*, vol. 22, no. 6, pp. 644–654, 1976.

[15] E. Franz and A. Pfitzmann, "Steganography secure against cover-stego-attacks," in *IH*, vol. 1768. Springer, 1999, pp. 29–46.

[16] M. Barni, F. Bartolini, and T. Furon, "A general framework for robust watermarking security," *Signal Processing*, vol. 83, no. 10, pp. 2069–2084, 2003.

[17] F. Cayre, C. Fontaine, and T. Furon, "Watermarking security: theory and practice," *IEEE TSP*, vol. 53, no. 10, pp. 3976–3987, 2005.

[18] S. Katzenbeisser and F. Petitcolas, "Defining security in steganographic systems," in *SWMC*, vol. 4675. SPIE, 2002, pp. 50–56.

[19] C. Ashokkumar, R. P. Giri, and B. Menezes, "Highly efficient algorithms for aes key retrieval in cache access attacks," in *EuroS&P*. IEEE, 2016, pp. 261–275.

[20] P. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *CRYPTO*. Springer, 1999, pp. 388–397.

[21] E. Cagli, "Feature extraction for side-channel attacks," Ph.D. dissertation, Sorbonne Université, 2018.

[22] O. Bronchain, "Worst-case side-channel security: from evaluation of countermeasures to new designs." Ph.D. dissertation, Catholic University of Louvain, 2022.

[23] A. Ker and T. Pevný, "Batch steganography in the real world," in *MMSec*. ACM, 2012, pp. 1–10.

[24] R. Cogranne, Q. Giboulot, and P. Bas, "Alaska#2: Challenging academic research on steganalysis with realistic images," in *WIFS*. IEEE, 2020, pp. 1–5.

[25] A. Mallet, M. Beneš, and R. Cogranne, "Cover-source mismatch in steganalysis: Systematic review," *EURASIP JIS*, 2024.

[26] J. Fridrich, M. Goljan, and R. Du, "Steganalysis based on JPEG compatibility," in *Multimedia Systems and Applications*, vol. 4518. SPIE, 2001, pp. 275–280.

[27] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE TIFS*, vol. 11, no. 2, pp. 221–234, 2015.

[28] M. Beneš, N. Hofer, and R. Böhme, "The effect of the JPEG implementation on the cover-source mismatch error in image steganalysis," in *EUSIPCO*. IEEE, 2022, pp. 1057–1061.

[29] T. Pevný, "Detecting messages of unknown length," in *MWSF*, vol. 7880. SPIE, 2011, pp. 300–311.

[30] P. Schöttle, "The role of side information in steganography," Ph.D. dissertation, Universität Münster, 2014.

[31] A. Ker, "The square root law requires a linear key," in *MMSec*. ACM, 2009, pp. 85–92.

[32] ——, "Locating steganographic payload via WS residuals," in *MMSec*. ACM, 2008, pp. 27–32.

[33] L. Pibre, P. Jérôme, D. Ienco, and M. Chaumont, "Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover source-mismatch," *arXiv*, 2015.

[34] M. Beneš, B. Lorch, and R. Böhme, "JPEG steganalysis using leaked cover thumbnails," in *WIFS*. IEEE, 2023, pp. 1–6.

[35] J. Fridrich, M. Goljan, and D. Hogea, "Steganalysis of JPEG images: Breaking the F5 algorithm," in *IH*. Springer, 2003, pp. 310–323.

[36] A. Ker, "Resampling and the detection of LSB matching in color bitmaps," in *SSWMC*, vol. 5681. SPIE, 2005, pp. 1–15.

[37] J. Kodovský and J. Fridrich, "Calibration revisited," in *MMSec*. ACM, 2009, pp. 63–74.

[38] M. Kirchner and R. Böhme, "Steganalysis in technicolor: Boosting WS detection of stego images from CFA-interpolated covers," in *ICASSP*. IEEE, 2014, pp. 3982–3986.

[39] M. Beneš and R. Böhme, "Exploring diffusion-inspired pixel predictors for WS steganalysis," in *IH&MMSec*. ACM, 2024, pp. 75–86.

[40] P. Bas, T. Filler, and T. Pevný, "Break our steganographic system: the ins and outs of organizing boss," in *IH*. Springer, 2011, pp. 59–70.

[41] E. Dworetzky, E. Kaziakhmedov, and J. Fridrich, "Improving steganographic security with source biasing," in *IH&MMSec*. ACM, 2024, pp. 19–30.

[42] V. Sedighi, J. Fridrich, and R. Cogranne, "Content-adaptive pentary steganography using the multivariate generalized gaussian cover model," in *MWSF*, vol. 9409. SPIE, 2015, pp. 144–156.

[43] J. Butora and P. Bas, "Side-informed steganography for JPEG images by modeling decompressed images," *IEEE TIFS*, vol. 18, pp. 2683–2695, 2023.

[44] A. Ker, "Batch steganography and pooled steganalysis," in *IH*. Springer, 2006, pp. 265–281.

[45] J. Fridrich and M. Goljan, "Robust hash functions for digital watermarking," in *InfoTech*. IEEE, 2000, pp. 178–183.

[46] A. Ker, "Steganographic strategies for a square distortion function," in *SFSWMC*, vol. 6819. SPIE, 2008, pp. 43–55.

[47] J. Fridrich, *Steganography in digital media: principles, algorithms, and applications*. Cambridge University Press, 2010.

[48] T.-V. Nakajima and A. Ker, "The syndrome-trellis sampler for generative steganography," in *WIFS*. IEEE, 2020, pp. 1–6.

[49] G. Kaptchuk, T. Jois, M. Green, and A. Rubin, "Meteor: Cryptographically secure steganography for realistic distributions," in *CCS*. ACM, 2021, pp. 1529–1548.

[50] R. Anderson, "Stretching the limits of steganography," in *IH*. Springer, 1996, pp. 39–48.

[51] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *WIFS*. IEEE, 2014, pp. 48–53.

[52] Q. Li, G. Feng, Y. Ren, and X. Zhang, "Embedding probability guided network for image steganalysis," *SPL*, vol. 28, pp. 1095–1099, 2021.

[53] P. Schöttle and R. Böhme, "Game theory and adaptive steganography," *IEEE TIFS*, vol. 11, no. 4, pp. 760–773, 2015.