# SoK: Kleptographic Attacks

Ting-Yun Yeh

Academia Sinica, Taipei 115201, Taiwan
yehtingyun0952@iis.sinica.edu.tw

**Abstract.** Kleptography was first proposed by Adam Young and Moti Yung in 1996, while algorithm substitution attack was introduced by Mihir Bellare et al. as a variation of kleptography in 2014 after the Dual EC incident with the confidential documents revelation by Edward Snowden. These two paradigms share a common goal: to enable attackers to embed covert capabilities into cryptographic implementations while maintaining the appearance of normal functionality. The goal of this paper is to consolidate existing research on kleptographic attacks, integrate it into a unified definition, and explore future directions for the research in this field. This paper begins by introducing and comparing the two major branches of kleptographic attacks: traditional kleptography and post-Snowden algorithm substitution attack, highlighting their theoretical distinctions, threat models, and historical development. Then, it analyzes the specific goals that attackers aim to achieve through such subversions and propose a generalized definition of algorithm substitution attack that include all the goals. The paper also presents practical examples framed within my definition and classify prior research works as either strong or weak attacks based on their structure and undetectability. Finally, it discusses the current landscape of research in kleptographic attacks, and then suggest future directions for the attack and defense perspectives.

**Keywords:** Kleptography · Algorithm Substitution Attack · Backdoor Cryptography · Subversion Attack · Cryptovirology

## 1 Introduction

The terminology *Kleptography* was firstly introduced in a paper published in CRYPTO'96, *The Dark Side of Black-Box Cryptography, or Should We Trust Capstone* by Adam Young and Moti Yung, which falls under the category of cryptovirology [28]. This attack model was inspired by the nature of cryptographic engineering involving black-box obfuscation, where algorithms are hidden to protect against exploitation of vulnerabilities or leakage of business secrets. However, when the internal workings of a cryptographic system are concealed, there is no guarantee that the system behaves correctly or securely. Is a black-box cryptographic design truly secure? Could it leak key information? Is it dangerous if a black-box system is reverse-engineered? Kleptography turns these concerns into practical threats by embedding malicious behavior within cryptographic implementations. Although the issue is critical, it has historically received limited attention in the cybersecurity and cryptography research communities. This

changed in 2013 when Edward Snowden revealed classified documents about global surveillance projects of the NSA. One notable example was the possible backdoor in the Dual EC, which used techniques consistent with kleptography, drawing widespread public and academic attention. In response, Mihir Bellare et al. proposed the concept of the *Algorithm Substitution Attack* (ASA) [4], a variation of kleptography that models large-scale subversions of encryption algorithms, particularly in the context of mass surveillance. A detailed comparison between kleptography and algorithm substitution attack is presented here. Moreover, as cryptography expands to emerging domains such as post-quantum and quantum cryptography, new research has explored how kleptographic attacks techniques can be adapted to these contexts.

The motivation of this paper is to serve as a foundational survey for a broader line of research on kleptographic attacks. This work aims to organize the body of work related to kleptography since its inception, provide a unified perspective, and lay the groundwork for future exploration. I hope this paper offers a clear and comprehensive entry point for new researchers who wish to explore this field.

This work contributes to the field of kleptographic attacks by proposing a unified and generalized definition that captures both traditional kleptography and algorithm substitution attack. It further introduces a classification framework and distinguishes between strong and weak forms of kleptographic attacks based on their structure and capabilities. In addition, it provides a comprehensive survey of existing literature, contextualizes recent developments, and suggests directions of future research in this evolving area.

### 1.1 Related Works

Young et al. developed the foundational structure of kleptography, beginning with their works [28,29,30,31,32,34,35,33]. Bellare et al. introduced the first formal model of algorithm substitution attacks in [4], commonly referred to as the BPR version. A subsequent variant, known as the DFP version, was proposed by Degabriele et al. in [10], which adapted the security model from the BPR framework. Later, Mihir Bellare et al. introduced a more powerful and general formulation of ASA, referred to as the BJK version, aiming to cover a broader class of symmetric encryption cryptosystems [3]. In the newest version of ASA [5], Bellare et al. further extended ASA into public-key cryptosystems.

Bernstein et al. investigated the entire Dual EC DRBG incident in [6], providing insights into how kleptographic techniques were potentially embedded into standardized cryptographic primitives. In a related study, Checkoway et al. analyzed the use of kleptographic attacks in the context of Dual EC DRBG and discussed its potential exploitation within TLS implementations [8].

The application of kleptographic attacks in digital signatures was demonstrated in [19,36,12], while their relevance to blockchain systems was explored in [1,24].

Recent studies have also investigated the use of kleptographic attacks in post-quantum cryptographic schemes, including Kyber [22,25,16,27] and (Classic) McEliece [15], demonstrating the applicability of such attacks to emerging cryptographic technologies.

In addition, several studies have focused on countermeasures against kleptographic attacks. Yevgeniy Dodis et al. introduced the concept of *immunizers*, special hash functions designed to neutralize trapdoors embedded in cryptosystems. [11] Mironov et al. proposed the notion of *reverse firewalls* as a defense mechanism against information leakage from inside [20]. Fischlin et al. also proposed countermeasures targeting ASA, aiming to detect and mitigate such attacks through self-guarding cryptographic protocols [14].

### 1.2 Structure of this Paper

- Section 2 provides a general introduction to kleptographic attacks with kleptography by Adam Young and Moti Yung and ASA by Mihir Bellare et al.
- Section 3 proposes a unified definition of kleptographic attacks and introduces a new framework for classifying strong and weak attacks.
- Section 4 analyzes with the kleptography on ElGamal and Dual EC, along with other existing works also be evaluated under the framework in this paper.
- Section 5 concludes the paper with observations and future directions for attack strategies and countermeasures.

## 2 Kleptographic Attacks

Kleptography is the study of how to subvert a cryptosystem by embedding malicious functionalities or backdoors while maintaining its apparent normal behavior.

To illustrate the intuition behind such attacks, let us consider a metaphor: suppose a cryptosystem is a lock on the door of a mansion. A malicious actor, who intends to gain unauthorized access, designs a new lock and convinces the owner to adopt it. This new lock functions normally but contains hidden mechanisms known only to the attacker, allowing him or her to open the door at will. Once installed, the attacker can silently observe the activities of the owners, steal valuables, or tamper with the interior, all without raising suspicion from the owners.

Back to the practice of cryptography, the structure of cryptographic systems is far more complex than door locks. Attackers may exploit their knowledge of cryptanalysis to compromise some components, such as encryption algorithms, key generation mechanisms, or random number generators.

To date, there are two major categories of kleptographic attacks are introduced.

This section introduces and compares both the categories. Specifically, here uses the term, kleptography (or klepto), to refer to the class of attacks originally introduced by Young et al., while the term, kleptographic attack, is used as a general term encompassing both kleptography and algorithm substitution attack discussed in this paper.

### 2.1 Kleptography

As mentioned in the introduction, kleptography is an attack designed to raise awareness of the potential dangers inherent in black-box cryptographic implementations. The core idea of kleptography is to subtly design or modify a cryptographic algorithm such that it contains a backdoor known only to the attacker. This allows the attacker to extract secret key information and compromise the system, all while preserving the appearance of correct functionality. Since black-box cryptography is often produced by industries, Adam Young et al. assumed in their original work that the attacker is the manufacturer [28].

**Black-Box Cryptography.** The terminology *black-box implementation* refers to a setting in which the internal structure of an algorithm is hidden. In the context of cryptography, a black-box implementation conceals the underlying details of the cryptographic algorithm, allowing users to interact with it only through its inputs and outputs. In contrast, a *white-box implementation* reveals its internal workings and logic to the user. In real-world, many widely-used cryptographic algorithms are standardized and publicly known. However, when industries implement these algorithms, they may choose to modify them to improve performance or protect their intellectual property, and release black-box versions. This approach offers merits for commerncial, but it is easily hideen vulnerabilities or intentional subversions simultaneously.

**SETUP Mechanism.** The Secretly Embedded Trapdoor with Universal Protection Mechanism, also known as the SETUP Mechanism, was introduced in the foundational work of Young et al. [28,29]. It defines how a kleptographic backdoor can be embedded in a black-box cryptosystem, enabling the attacker to recover secret inforamtion while preserving public correctness and indistinguishability. The definition of SETUP is given in the following [28,29]:

**Definition 1.** *Assume that $S$ is a black-box cryptosystem with a publicly known specification. A (regular) SETUP mechanism is an algorithmic modification made to $S$ to get $S'$ such that:*

1. *The input of $S'$ agrees with the public specifications of the input of $S$.*
2. *$S'$ computes efficiently using the attacker's public encryption function $E$ (and possibly other functions as well), contained within $S'$.*
3. *The attacker's private decryption function $D$ is not contained within $S'$ and is known only by the attacker.*

4. *The output of $S'$ agrees with the public specifications of the output of $S'$. At the same time, it contains published bits (of the user's secret key) which are easily derivable by the attacker (the output can be generated during key-generation or during system operation like message sending).*
5. *Furthermore, the output of $S$ and $S'$ are polynomially indistinguishable to everyone except the attacker.*
6. *After the discovery of the specifics of the setup algorithm and after discovering its presence in the implementation (e.g. reverse engineering of hardware tamper-proof device), users (except the attacker) cannot determine past (or future) keys. (This is an extra definition in [29].)*

To preserve the illusion of normality, the interaction between the system and users must remain consistent. Item *1.* ensures the input space of $S'$ aligns with the input space of $S$, while item *4.* requires that the output space remains consistent with expectations. To avoid detection of subversions, item *5.* and item *6.* require that the outputs from $S$ to $S'$ are polynomially indistinguishable, which means no efficient computation can distinguish between them within polynomial time. This also implies that even if the system is reverse-engineered, users cannot deduce how the tampered algorithm generates or leaks secret information. It is also important to note that item *3.* highlights the asymmetry of decryption functions from the users to the attacker. The user operates with a standard decryption function, while the attacker possesses an additional decryption function which allows him or her to extract secret information, mostly the keys, from the embedded backdoor. This attack typically involves modifying the key generation algorithm. Moreover, the original SETUP definition implicitly refines to some restrictions. Kleptography is generally applicable to asymmetric cryptosystems, and the subversions commonly target key generation mechanisms.

Furthermore, Adam Young and Moti Yung further categorized SETUP Mechanisms into Weak SETUP and Strong SETUP [29].

**Definition 2.** *A weak SETUP is a regular SETUP except that the output of $S$ and $S'$ are polynomially indistinguishable to everyone except the attacker and the owner of the device who is in control of his or her own private key. A strong SETUP is a regular setup, but in addition, to assume that the users are able to reversely engineer the device, they are able to analyze the actual implementation of $S'$ and deploy the device. However, ther users still cannot recover the previous and future keys, and if the SETUP is not always applied to future keys, then $S$ and $S'$ still remain polynomially indistinguishable.*

The difference between the weak, regular, and strong SETUP Mechanism is based on the distinguishability between $S$ and $S'$.

### 2.2 Algorithm Substitution Attack

The Algorithm Substitution Attack (ASA) is the other type of kleptographic attacks, adapted from the original model by Mihir Bellare, Kenneth G. Paterson, and Phillip Rogaway in response to the incident of Dual EC. Their threat

model targets mass-surveillance carried out by large-scale organizations such as governments. While kleptography focuses on implementation level subversion, ASA is formulated using game-based security definitions to analyze the intensity of the attacks. ASA was originally developed for symmetric-key encryption systems, where the attacker subverts the encryption algorithm to embed a covert channel. In some cases, the attacker may also inject a maliciously crafted initial nonce to further control the outputs.

More recently, Mihir Bellare, Doreen Riepel, and Laura Shea extended the ASA model to public algorithm [5]. In this variation, the attacker targets components such as hash functions or signature verification algorithms rather than the encryption schemes. Moreover, the target of this attack is a public algorithm where no secret key to exfiltrate. This is the main different between public algorithm ASA (PA-SA) and the original ASA (secret algorithm ASA or SA-SA).

**(Un)detectability.** An important security notion introduced with ASA is *detectability*, which quantifies the likelihood how a user can distinguish between a normal or a subverted implementation. To execute surveillance, attackers aim to remain undetected; therefore, the attack must be stealthy even under observation. This particularly important in standardized cryptographic systems such as Dual EC, where the algorithm is publicly scrutinized. Thus, the definition of detectability assumes that the users are still unable to distinguish whether the system is a normal or a subverted algorithm even though they have the knowledge of the algorithm. Detectability for SA-SA is given in the following [4]:

**Definition 3.** *Let $\Pi = (\mathcal{K}, \mathcal{E}, \mathcal{D})$ be an original cryptosystem, and $\Pi' = (\mathcal{K}', \mathcal{E}', \mathcal{D}')$ be the subverted cryptosystem. Let $\mathcal{U}$ be an algorithm of a detection test run by users. Then, the advantage of detectability is defined as*

$$Adv^{det}_{\Pi,\Pi'}(\mathcal{U}) = 2Pr[DETECT^{\mathcal{U}}_{\Pi,\Pi'} \Rightarrow true] - 1$$

This measures the probability of the detector $\mathcal{U}$ in distinguishing whether the cipher is produced by $\Pi$ or $\Pi'$. If the value is small enough to be negligible, then the attack is hardly detectable.

The corresponding security game is defined in Fig. 1.

Here, $A$ denotes is associated data, $\sigma_i$ denotes the state, and $\varepsilon$ represents the empty string.

In PA-SA, the general idea is similar, but the encryption algorithms $\mathcal{E}$ and $\mathcal{E}'$ are replaced with hash or verification functions $F$ and $F'$. The modified security definition for PA-SA is defined as [5]:

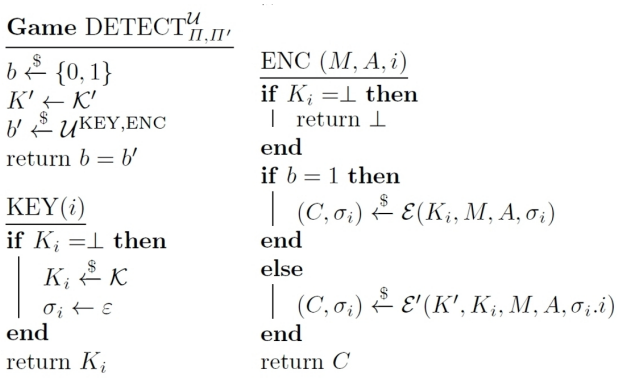


**Fig. 1.** Security game of SA-SA detectability [4].

**Definition 4.** *Let $F$ be the original algorithm and $F'$ be the subverted algorithm, while $(U)$ is an algorithm of a detection test run by the users. Then, the advantage of detectability is defined as*

$$Adv^{det}_{F,F'}(\mathcal{U}) = 2Pr[DETECT^{\mathcal{U}}_{F,F'} \Rightarrow true] - 1$$

The corresponding security game is shown in Fig. 2.

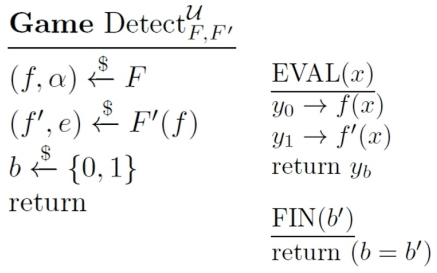


**Fig. 2.** Security game of PA-SA detectability [5].

The security game for PA-SA begins by initializing the original algorithm $F$, which generates a function $f$ along with auxiliary information $\alpha$. Then, the subversion algorithm $F'$ takes $f$ as input and produces a subverted version $f'$ along with an exploitation algorithm $e$, which is kept secret by the attacker. The oracle EVAL($x$) returns either $f(x)$ or $f'(x)$ depending on a guessing bit $b$.

To provide a clear overview, we summarize the comparison between the two

| | SA-SA | PA-SA |
|---|---|---|
| state | no restriction but most are black-box [3] | black-box |
| cryptosystem | symmetric encryption | public algorithms |
| target parts | encryption algorithm | hash function or verification |
| exploitation | covert in ciphers or control initial vector | triggered by auxiliary information |
| attacker's goal | exfiltrate messages | passing authentication |
| impact | decrypt users' communication | forge signatures and certification |

**Table 1.** Comparison between SA-SA and PA-SA.

ASAs in Table 1.

## 3 A New Definition

To provide a unified perspective on kleptographic attacks, here proposes an improved definition and threat model that integrates the ideas from both the original kleptography [28] and algorithm substitution attacks [4]. We also present a classification framework to distinguish between strong and weak kleptographic attacks.

### 3.1 A New SETUP

We first consider who might want to break into the mansion. Potential malicious actors could include burglars aiming to steal valuables, governments conducting surveillance, or malicious actors seeking to sabotage specific targets. Since our definition of kleptographic attacks aims to be more general, we make no assumptions about the identity of attackers. The common goal for all of the models is to surreptitiously subvert a cryptosystem. To ensure the malicious lock appears legitimate, the cryptosystem must preserve its ability to encrypt and decrypt correctly. Decryption failures would be obviously detected by users. At the same time, the attacker must be able to derive useful information from the outputs. Therefore, although the system must remain operational, it cannot guarantee that ciphertexts or decrypted messages behave exactly as users expect. We also revise the assumption that cryptosystems are always black-box implemented. In modern practice, since many cryptosystems are open-source or standardized, algorithmic opacity is no longer a necessary condition.

To summarize, we formalize our new SETUP Mechanism below.

**Definition 5.** *Let $S = (\mathcal{K}, \mathcal{E}, \mathcal{D})$ be a regular cryptosystem and $S' = (\mathcal{K}', \mathcal{E}', \mathcal{D}')$ be a subverted version of $S$. A new SETUP Mechanism is an algorithmic modification from $S$ to $S'$ such that:*

1. *The input space of $S'$ agrees with that of $S$, i.e., both $S$ and $S'$ share the same message space $\mathcal{M}$.*
2. *The ciphertext space $\mathcal{C}'$ of $S'$ is a subset of the ciphertext space $\mathcal{C}$ of $S$.*
3. *The subversion from $S$ to $S'$ is undetectable, even if the design of $S'$ is fully disclosed.*
4. *The subverted components may include one or more parts of the cryptosystem.*
5. *The attacker can either partially or fully control the output of $S'$, or at least predict its pattern.*

To clearify the overview, we provide the comparison among kleptography, symmetric and public-key ASA, and the new framework of kleptographic attack in Table 2.

| | kleptography | SA-SA | PA-SA | this paper |
|---|---|---|---|---|
| state | black-box | no restriction but most are black-box | black-box | no restriction |
| cryptosystem | no specific but mainly asymmetric | symmetric | public algorithms | no restriction |
| target parts | key generation or key exchange | encryption algorithm | hash function or verification | any part(s) of cryptosystem |
| exploitation | covert secrets in key generations so that attackers can recover later | covert in ciphers or control initial vector | triggered by auxiliary information | algorithmic subversion |
| malicious actors | manufacturers | governments or large-scale organizations | governments or large-scale organizations | manufacturers, governments, spies, criminals, etc. |
| attacker's goal | exfiltrate keys | exfiltrate messages | passing authentication | get access to users' systems |
| impacts | steal the secrets from users | decrypt users' communication to reach surveilance | forge authentication to reach surveillance | various and the previous all are included |

**Table 2.** Comparison among kleptography, the two ASAs, and this work.

### 3.2 Intensity of Kleptographic Attack

In kleptography, attack strength is distinguished via Strong and Weak SETUP. Strong SETUP grants the attacker greater control and makes it significantly harder for users to predict future key generation patterns. The difficulty of distinguishing between $S$ and $S'$ is a major factor in determining attack intensity. On the other hand, algorithm substitution attacks do not explicitly label attacks as strong or weak. Instead, its game-based definition of detectability provides a quantitative measurement. The less of the value of advantage, the harder it is to detect, and the stronger the attack. Reversely, the larger of the value of advantage leads to the weaker the attack. Returning again to the mansion-lock metaphor, our objective as the attacker is to gain entry. We do so by convincing the owner to install our tampered lock. A successful attack means we can access the mansion unnoticed. Failure can result from:

- The backboor being detected and removed. We must ensure undetectability.
- The victim refusing to adopt our lock. We must increase incentives, such as improved efficiency.

In both kleptography and ASA, the core metric for evaluating attack success is indistinguishability or undetectability. To formalize a boundary between strong and weak attacks, we borrow the standard cryptographic notion of negligible advantage [2]. If the advantage of adversary in detector is $\leq 2^{-p}$ for a $p$-bit security parameter, we classify the attack as a strong kleptographic attack. If the advantage lies between $2^{-p}$ and $2^{-p/2}$, it is considered a weak kleptographic attack. In practice, a reduction in effective security from $p$ to $p/2$ is still considered secure for most applications. Hence, detectability with advantage less than $2^{-p/2}$ is still acceptable in our model.

On the other hand, the second point regarding user adoption incentives does not apply uniformly to all kleptographic attacks. This distinction becomes clearer when we consider two different scenarios. In the first case, the attacker subverts an existing standardized cryptosystem. Since the algorithm is already well-known and widely used, claiming significant performance improvements or modifications might raise suspicion from experts or the community. Subtle changes could be scrutinized more closely, making it harder for the attacker to inject a backboor without detection. In contrast, if the attacker designs an entirely new cryptographic algorithm from scratch, such as the case of Dual EC [6], they have more flexibility in controlling the narrative. By emphasizing certain desirable properties like provable security, compliance with standards, or high efficiency, the attacker can more easily persuade users or instittions to adopt the new cryptosystem, despite the presence of a hidden backdoor. Thus, while performance incentives may be ineffective or even counterproductive in the case of subverting known systems, they can play a crucial role in the success of new, maliciously designed cryptosystems.

## 4 Cases of Kleptographic Attacks

To better understand how a kleptographic attack manifests in real-world scenarios, this section presents and analyzes several representative examples. These case studies highlight the diversity of attack vectors and their conformity with the proposed framework, including the intensity classifications introduced in Section 3.

### 4.1 Case 1: Klepto on ElGamal

This is one of the earliest example of kleptography given in [28], targeting the ElGamal public-key encryption scheme. Here revisits the case to examine its structure under the new framework. The original algorithm of ElGamal encryption is as following [13]:

- Public Key: $p$, $g$, $y$, $h$
- Private Key: $x$
- Key Generation: in a finite generated group $G$ with $p$ elements, randomly pick elements $g$ and $x$ to compute $h = g^x \mod p$
- Encryption:
  - Randomly select an element $y$ from $G$ and compute $c_1 = g^y \mod p$.
  - Generate a share secret $s = h^y \mod p$.
  - Compute the cipher $c_2 = m \cdot s \mod p$.
  - Return the pair $c = (c_1, c_2)$.
- Decryption:
  - Take the private key to recover $s = c_1^x = (g^y)^x \mod p$.
  - Recover the message $m = c_2 \cdot s^{-1} \mod p$.

In the kleptography provided in [28], the attacker attempts to subvert the key generation algorithm. The subversion processes as follows:

1. Randomly select $y$.
2. If $gcd(y, p-1) \neq 1$ is coprime to $p-1$, jump to Step 1.
3. Compute $a = x \cdot h^y \mod p$, where $x$ is the private key.
4. Encrypt $a$ again with another private key $sk \in G$ which is generated by attacker.
5. If $a$ (after encryptoin) is not prime $\vee$a $\geq x$, jump to Step 4.
6. Compute $b = g^y \mod p$.
7. If $a \geq k$, jump to Step 1.
8. Set $p = a$, $g = b$.
9. Return $(g, p, y)$ as the public key to users

The attacker can use the information of public key to recover the user private key $x$. By decrypting $p$ and $g$, we get $x = p/g^{sk} = a/b^{sk} = x \cdot (g^{sk})^y/(g^y)^{sk} \mod p$.

Under the new definition of kleptographic attacks, it first requires the subverted

cryptosystem to accept the same message space from the original cryptosystem, and the cipher space cannot be more than the original one. Since the subverted ElGamal encryption only contaminated the key generation, and the restrictions stated in the kleptography such like picking $y$ coprime to $p-1$, here can see the decrement of key space would not let the cipher space more than the original. Besides, the attacker is able to get the private key from public key information, this also passes the last item of definition. In the detectability test, the decrement of the key space is exploited to be the strategies for detectors. The attack above restricted selection of $y$ being coprime to $p-1$. In the best case, we suppose $p-1$ is an exponent with base 2; therefore, the decrement of the key space is at least half or one bit less of security level. This falls into the category of the **weak kleptographic attack**.

### 4.2 Case 2: Dual EC

Dual EC was standardized by the National Institute of Standards and Technology (NIST) in 2006 as SP 800-90 [17,6]. From the outset, it was met with skepticism by the cryptographic community, despite its widespread industrial adoption at the time. Until the incident of Snowden, the existence of backdoor in Dual EC was eventually confirmed. After the incident, the work of Stephen Checkoway et al. demonstrated that Dual EC was fundamentally insecure, and could allow malicious actors to compromise SSL/TLS communications [8]. In a separate incident, it was discovered that Dual EC introduced serious vulnerabilities in Juniper Network's VPN implementation, further highlighting its potential for exploitation and causing widespread distrust within the industry [7].

The full name of Dual EC is the Dual Elliptic Curve Deterministic Random Bit Generator (Dual EC DRBG), a pseudorandom number generator (PRNG) based on elliptic curve operations. It operates using two fixed points, $P$ and $Q$, on the NIST P-256 elliptic curve, and a 256-bit internal state $s$. A function $f$ is defined as $f(s_0) = s_1 = x(s_0P)$, which multiplies the state $s_0$ with point $P$, and returns the $x$-coordinate of the resulting point as the new state $s_1$. A second function $g$ then performs an analogous operation using point $Q$: $g(s_1) = x(s_1Q) = r_1$. The final output is derived from the most significant 16 bytes of $r_1$, with a maximum output length of 30 bytes. If the requested output is not a multiple of 30, the remaining bits are truncated from the most significant end. For each subsequent output, the internal state is updated to $s_1$ and the process is repeated [6,8]. A schematic representation of this process is shown in Fig.3.

Dan Shumow and Niels Ferguson proposed an attack on Dual EC assuming the attacker knows a scalar $d$ such that $P = dQ$ [9]. Given this assumption and the observed output $r_1$, the attacker who knows the curve parameters can reconstruct the full point $R_1 = (r_1, y_{r_1})$. By multiplying both sides of this point with the scalar $d$, the attacker computes:

$$d(r_1, y_{r_1}) = dR_1 = ds_1Q = s_1dQ = s_1P$$

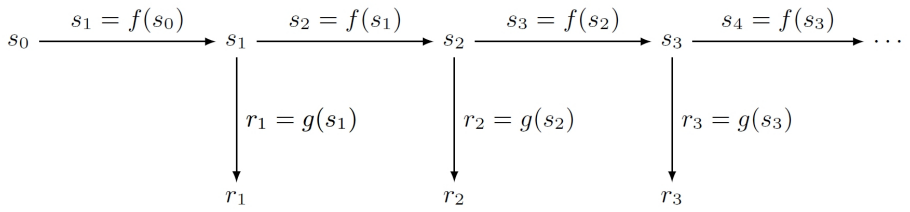


**Fig. 3.** The schematic of the algorithm of Dual EC [6].

thereby recovering $s_1P$ and thus learning the internal state. Once the attacker knows $s_1$, all future states and outputs of the PRNG can be predicted. This attack is made feasible by the fact that the standard does not specify how $Q$ is generated and both $P$ and $Q$ are published constants. Therefore, the possibility that $P$ and $Q$ are related by a known scalar $d$ constitutes a backdoor, and hence a form of kleptographic attack. That is to said, the link between $P$ and $Q$ is based on the discrete logarithm problem, which is computationally hard.

Putting aside the controversy surrounding whether the backdoor was intentional, the design still suffers from inherent structural weaknesses. In the framework of kleptographic attacks, it typically considers an entire cryptosystem as the unit of analysis. While a PRNG is not a complete cryptosystem on its own, it plays a foundational role in generating random inputs for key generation, encryption, and other cryptographic operations. Therefore, if a subverted PRNG embedded into a larger cryptographic system, it can serve as a vector for kleptographic subversion. In this light, although Dual EC alone is not a standalone kleptographic attack, it can nonetheless contaminate a larger system and effectively enable one.

According to the new definition of kleptographic attacks, we require an "original" and a "subverted" cryptosystem for comparison. In this case, here adopts the interpretation from Shumow and Ferguson's attack, where the subverted version uses fixed $P$ and $Q$, while the original uses randomly generated $Q$ per instance [9]. First, both versions accept the same message space, satisfying the first condition. Second, since the subverted version has fixed parameters, its output space is a strict subset of the original's. As shown earlier, the attacker can derive the internal state from the output using knowledge of the scalar $d$, enabling prediction of future outputs. In the detectability game, here allows the detector to guess the scalar $d$ relating $P$ and $Q$. However, since the space of $d$ is of size $2^{256}$, it makes detection practically infeasible. This meets the threshold of being strongly undetectable under the proposed classification. In conclusion, Dual EC constitutes a **strong kleptographic attack** when embedded within a cryptosystem.

### 4.3 Other Examples

To further demonstrate the applicability of the framework, here categorizes additional known attacks from the literature in Table 3.

| Kleptographic Attacks List | | |
|---|---|---|
| Strong Attacks | Weak Attacks | Non Attacks |
| Dual EC<br>[12]<br>[25]<br>[27] | Klepto on ElGamal<br>[22] | [15]<br>[16]-1, [16]-2, [16]-3<br>[18]<br>[21] |

**Table 3.** categorization of attacks from related literatures

Addition notes:

- [12]: uses an implementation-level subversion unlike other instances using mathematical subversion to trigger a hidden attacker mode via a deliberate bug. Its detectability is low, qualifying it as strong.
- [25], [27]: correspond with the new framework and detectability is negligible to be defined as strong attacks.
- [22]: introduces minor decryption failure riks due to contaminated key generation. However, this can be exploited by detectors, making it a weak kleptographic attack.
- [15], [16] (3 attacks in this work), [21]: do not meet the undetectability threshold and are classified as non-kleptographic.
- [18]: intentionally causes decryption failure, violating the requirement that subverted cryptosystems remain functionally correct, and thus falls outside the scope of kleptographic attacks.

These examples collectively illustrate the applicability and robustness of the new framework in classifying and analyzing kleptographic attacks across a variety of cryptosystems.

## 5 Conclusion

Kleptographic attacks represent a significant and insidious threat in modern cryptography. As demonstrated through the case studies in this paper, both strong and weak variants of these attacks remain difficult to detect, even with sophisticated detection strategies. In the detectability framework, successful mitigation often relies on cryptographic knowledge of defenders. A poorly chosen strategy may not only fail to detect the attack but could also lead to a false sense of security. Furthermore, many end users or even system integrators lack

the technical background necessary to defend against such advanced threats, leaving them especially vulnerable. Even in situations where a potential backdoor is discovered, conclusively identifying the subversion as a kleptographic attack remains a complex challenge. For example, in the case of Dual EC, the cryptographic community widely suspected a hidden backdoor, yet attackers could still plausibly claim that the design flaws were unintentional.

Nevertheless, historical evidence indicates that kleptographic attacks are especially suitable for use by state-level actors. Governments can use them to covertly monitor individuals, particularly political dissidents or ideological opponents. Beyond domestic surveillance, kleptographic subversion serves as a powerful tool for cyber-espionage. Malicious backdoors planted in exported cryptographic products can be used to spy on foreign governments or industries, all while remaining undetected.

While the analysis emphasizes nation-state actors, it is important to note that these attacks are not necessarily exclusive to them. However, launching such attacks often requires privileged access to cryptographic design or standardization processes, something that is typically within the reach of governments or well-funded organizations. Fortunately, the open and standardized nature of most modern cryptographic algorithms ensures that security researchers can scrutinize these designs. When anomalies are detected, the cryptographic community is often quick to raise awareness. As a result, companies that insist on deploying flawed algorithms must either justify their decisions to a skeptical public or compensate for the associated reputational and economic costs.

### 5.1 Countermeasures

Several countermeasures have been proposed to mitigate the risk of kleptographic subversions.

**Immunizers.** Proposed by Dodis et al. in [11], immunizers transform the outputs of PRNGs in such a way that the attacker loses the ability to exploit any subverted behavior. These post-processing functions aim to neutralize the structure that enables backdoor exploitation, thereby improving the unpredictability and security of the output.

**Reverse Firewalls.** Introduced by Mironov et al. in [20], reverse firewalls act as a proxy layer that monitors and sanitizes data sent to and from potentially compromised cryptographic implementations. Unlike traditional firewalls, which defend against external threats, reverse firewalls protect against data exfiltration by modifying messages on the fly to prevent attackers from decoding embedded secrets.

**Detectors.** Detectors attempt to identify anomalies in cryptographic behavior solely by analyzing inputs and outputs, without requiring access to the internal algorithm. In some cases, detectors can also actively break the contaminated functionality, rendering the attack ineffective. One notable example is the Watchdog system [23], which introduces statistical analysis and challenge-response protocols to expose subversion.

**Self-guarding Cryptographic Protocols.** Fischlin et al. [14] proposed a class of cryptographic protocols designed to defend against kleptographic attacks intrinsically. These protocols assume that the initialization phase is executed on an honest implementation, during which it collects baseline samples such as ciphertexts or signatures. In later phases, even if the implementation becomes subverted, the protocol can still operate securely using the collected data without leaking sensitive information.

### 5.2 Further Research

While this paper introduces a new framework for defining and analyzing kleptographic attacks, several important research directions remain open for exploration:

*Kleptographic Attacks on Post-Quantum Cryptosystem.* Post-quantum cryptosystems are an emerging field, and many practitioners are still unfamiliar with its structural properties. This unfamiliarity creates an opportunity for attackers to design novel subversions. Although some initial research has examined subversion on NTRU [18], Kyber [26,27,22,25,16], and Classic McEliece [15], further work is needed to assess the vulnerability of other standaridized schemes such as HQC, SPHINCS+, and Falcon.

*Implementation of Kleptographic Attacks in Real-World Cryptography.* Although Dual EC remains the most renowned real-world example of a kleptographic attack, other domains are equally at risk. Network protocols such like TLS or VPN, blockchain infrastructure, cryptographic hardware, and embedded systems in critical applications such as autonomous vehicles or drones) are all potential targets. Realistic implementations and demonstrations of kleptographic subversions in these contexts are essential to raising awareness and improving defenses.

*Detecting Kleptographic Attacks by Formal Methods or Side-Channel Analysis.* Kleptographic attacks preserve functionality from a black-box perspective but often violate internal implementation correctness. This discrepancy opens the door to detection through formal methods or side-channel analysis. For instance, subverted algorithms may introduce additional operations in key generation or encryption. These modifications could be revealed through timing analysis, power consumption, or etc. Exploring such techniques may offer new hope for reliable detection in practice.

## References


1. Alsalami, N., Zhang, B.: Uncontrolled randomness in blockchains: Covert bulletin board for illicit activity. Cryptology ePrint Archive, Report 2018/1184 (2018), https://eprint.iacr.org/2018/1184
2. Bellare, M.: A note on negligible functions. Journal of Cryptology **15**(4), 271–284 (Sep 2002). https://doi.org/10.1007/s00145-002-0116-x
3. Bellare, M., Jaeger, J., Kane, D.: Mass-surveillance without the state: Strongly undetectable algorithm-substitution attacks. In: Ray, I., Li, N., Kruegel, C. (eds.) ACM CCS 2015: 22nd Conference on Computer and Communications Security. pp. 1431–1440. ACM Press, Denver, CO, USA (Oct 12–16, 2015). https://doi.org/10.1145/2810103.2813681
4. Bellare, M., Paterson, K.G., Rogaway, P.: Security of symmetric encryption against mass surveillance. In: Garay, J.A., Gennaro, R. (eds.) Advances in Cryptology – CRYPTO 2014, Part I. Lecture Notes in Computer Science, vol. 8616, pp. 1–19. Springer Berlin Heidelberg, Germany, Santa Barbara, CA, USA (Aug 17–21, 2014). https://doi.org/10.1007/978-3-662-44371-2_1
5. Bellare, M., Riepel, D., Shea, L.: Algorithm substitution attacks on public functions. Cryptology ePrint Archive, Report 2024/536 (2024), https://eprint.iacr.org/2024/536
6. Bernstein, D.J., Lange, T., Niederhagen, R.: Dual ec: A standardized back door. In: LNCS Essays on The New Codebreakers - Volume 9100. p. 256–281. Springer-Verlag, Berlin, Heidelberg (2015). https://doi.org/10.1007/978-3-662-49301-4_17, https://doi.org/10.1007/978-3-662-49301-4_17
7. Checkoway, S., Maskiewicz, J., Garman, C., Fried, J., Cohney, S., Green, M., Heninger, N., Weinmann, R.P., Rescorla, E., Shacham, H.: A systematic analysis of the juniper dual EC incident. In: Weippl, E.R., Katzenbeisser, S., Kruegel, C., Myers, A.C., Halevi, S. (eds.) ACM CCS 2016: 23rd Conference on Computer and Communications Security. pp. 468–479. ACM Press, Vienna, Austria (Oct 24–28, 2016). https://doi.org/10.1145/2976749.2978395
8. Checkoway, S., Niederhagen, R., Everspaugh, A., Green, M., Lange, T., Ristenpart, T., Bernstein, D.J., Maskiewicz, J., Shacham, H., Fredrikson, M.: On the practical exploitability of dual EC in TLS implementations. In: Fu, K., Jung, J. (eds.) USENIX Security 2014: 23rd USENIX Security Symposium. pp. 319–335. USENIX Association, San Diego, CA, USA (Aug 20–22, 2014)
9. Dan Shumow, N.F.: On the possibility of a back door in the nist sp800-90 dual ec prng. CRYPTO 2007 Rump Session (August 2007), https://rump2007.cr.yp.to/15-shumow.pdf
10. Degabriele, J.P., Farshim, P., Poettering, B.: A more cautious approach to security against mass surveillance. In: Leander, G. (ed.) Fast Software Encryption – FSE 2015. Lecture Notes in Computer Science, vol. 9054, pp. 579–598. Springer Berlin Heidelberg, Germany, Istanbul, Turkey (Mar 8–11, 2015). https://doi.org/10.1007/978-3-662-48116-5_28
11. Dodis, Y., Ganesh, C., Golovnev, A., Juels, A., Ristenpart, T.: A formal treatment of backdoored pseudorandom generators. In: Oswald, E., Fischlin, M. (eds.) Advances in Cryptology – EUROCRYPT 2015, Part I. Lecture Notes in Computer Science, vol. 9056, pp. 101–126. Springer Berlin Heidelberg, Germany, Sofia, Bulgaria (Apr 26–30, 2015). https://doi.org/10.1007/978-3-662-46800-5_5
12. Dubois, R.: Trapping ECC with invalid curve bug attacks. Cryptology ePrint Archive, Report 2017/554 (2017), https://eprint.iacr.org/2017/554

13. Elgamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory **31**(4), 469–472 (1985). https://doi.org/10.1109/TIT.1985.1057074
14. Fischlin, M., Mazaheri, S.: Self-guarding cryptographic protocols against algorithm substitution attacks. In: Chong, S., Delaune, S. (eds.) CSF 2018: IEEE 31st Computer Security Foundations Symposium. pp. 76–90. IEEE Computer Society Press, Oxford, UK (Jul 9–12, 2018). https://doi.org/10.1109/CSF.2018.00013
15. Hemmert, T., May, A., Mittmann, J., Schneider, C.R.T.: How to backdoor (classic) McEliece and how to guard against backdoors. In: Cheon, J.H., Johansson, T. (eds.) Post-Quantum Cryptography - 13th International Workshop, PQCrypto 2022. pp. 24–44. Springer, Cham, Switzerland, Virtual Event (Sep 28–30, 2022). https://doi.org/10.1007/978-3-031-17234-2_2
16. Joux, A., Loss, J., Wagner, B.: Kleptographic attacks against implicit rejection. Cryptology ePrint Archive, Report 2024/260 (2024), https://eprint.iacr.org/2024/260
17. Kelsey, J.: Dual ec in x9.82 and sp 800-90. NIST (May 2014), https://csrc.nist.gov/csrc/media/projects/crypto-standards-development-process/documents/dualec_in_x982_and_sp800-90.pdf
18. Kwant, R., Lange, T., Thissen, K.: Lattice klepto - turning post-quantum crypto against itself. In: Adams, C., Camenisch, J. (eds.) SAC 2017: 24th Annual International Workshop on Selected Areas in Cryptography. Lecture Notes in Computer Science, vol. 10719, pp. 336–354. Springer, Cham, Switzerland, Ottawa, ON, Canada (Aug 16–18, 2017). https://doi.org/10.1007/978-3-319-72565-9_17
19. Liu, C., Chen, R., Wang, Y., Wang, Y.: Asymmetric subversion attacks on signature schemes. In: Susilo, W., Yang, G. (eds.) ACISP 18: 23rd Australasian Conference on Information Security and Privacy. Lecture Notes in Computer Science, vol. 10946, pp. 376–395. Springer, Cham, Switzerland, Wollongong, NSW, Australia (Jul 11–13, 2018). https://doi.org/10.1007/978-3-319-93638-3_22
20. Mironov, I., Stephens-Davidowitz, N.: Cryptographic reverse firewalls. In: Oswald, E., Fischlin, M. (eds.) Advances in Cryptology – EUROCRYPT 2015, Part II. Lecture Notes in Computer Science, vol. 9057, pp. 657–686. Springer Berlin Heidelberg, Germany, Sofia, Bulgaria (Apr 26–30, 2015). https://doi.org/10.1007/978-3-662-46803-6_22
21. Posteuca, R., Ashur, T.: How to backdoor a cipher. Cryptology ePrint Archive, Report 2021/442 (2021), https://eprint.iacr.org/2021/442
22. Ravi, P., Bhasin, S., Chattopadhyay, A., Aikata, Roy, S.S.: Backdooring post-quantum cryptography: Kleptographic attacks on lattice-based KEMs. Cryptology ePrint Archive, Report 2022/1681 (2022), https://eprint.iacr.org/2022/1681
23. Russell, A., Tang, Q., Yung, M., Zhou, H.S.: Cliptography: Clipping the power of kleptographic attacks. In: Cheon, J.H., Takagi, T. (eds.) Advances in Cryptology – ASIACRYPT 2016, Part II. Lecture Notes in Computer Science, vol. 10032, pp. 34–64. Springer Berlin Heidelberg, Germany, Hanoi, Vietnam (Dec 4–8, 2016). https://doi.org/10.1007/978-3-662-53890-6_2
24. Tiwari, P.R., Green, M.: Algorithm-substitution attacks on cryptographic puzzles. Cryptology ePrint Archive, Report 2022/477 (2022), https://eprint.iacr.org/2022/477
25. Xia, W., Wang, G., Gu, D.: Post-quantum backdoor fornbsp;kyber-kem. In: Selected Areas in Cryptography – SAC 2024: 31st International Conference, Montreal, QC, Canada, August 28–30, 2024, Revised Selected Papers, Part I. p. 237–255. Springer-Verlag, Berlin, Heidelberg (2025). https://doi.org/10.1007/978-3-031-82852-2_11, https://doi.org/10.1007/978-3-031-82852-2_11

26. Yang, Z., Xie, T., Pan, Y.: Lattice klepto revisited. In: Sun, H.M., Shieh, S.P., Gu, G., Ateniese, G. (eds.) ASIACCS 20: 15th ACM Symposium on Information, Computer and Communications Security. pp. 867–873. ACM Press, Taipei, Taiwan (Oct 5–9, 2020). https://doi.org/10.1145/3320269.3384768
27. Yang, Z., Chen, R., Li, C., Qu, L., Yang, G.: On the security of lwe cryptosystem against subversion attacks. The Computer Journal **63**(1), 495–507 (2020). https://doi.org/10.1093/comjnl/bxz084
28. Young, A., Yung, M.: The dark side of "black-box" cryptography, or: Should we trust capstone? In: Koblitz, N. (ed.) Advances in Cryptology – CRYPTO'96. Lecture Notes in Computer Science, vol. 1109, pp. 89–103. Springer Berlin Heidelberg, Germany, Santa Barbara, CA, USA (Aug 18–22, 1996). https://doi.org/10.1007/3-540-68697-5_8
29. Young, A., Yung, M.: Kleptography: Using cryptography against cryptography. In: Fumy, W. (ed.) Advances in Cryptology – EUROCRYPT'97. Lecture Notes in Computer Science, vol. 1233, pp. 62–74. Springer Berlin Heidelberg, Germany, Konstanz, Germany (May 11–15, 1997). https://doi.org/10.1007/3-540-69053-0_6
30. Young, A., Yung, M.: The prevalence of kleptographic attacks on discrete-log based cryptosystems. In: Kaliski, Jr., B.S. (ed.) Advances in Cryptology – CRYPTO'97. Lecture Notes in Computer Science, vol. 1294, pp. 264–276. Springer Berlin Heidelberg, Germany, Santa Barbara, CA, USA (Aug 17–21, 1997). https://doi.org/10.1007/BFb0052241
31. Young, A., Yung, M.: Monkey: Black-Box symmetric ciphers designed for MONopolizing KEYs. In: Vaudenay, S. (ed.) Fast Software Encryption – FSE'98. Lecture Notes in Computer Science, vol. 1372, pp. 122–133. Springer Berlin Heidelberg, Germany, Paris, France (Mar 23–25, 1998). https://doi.org/10.1007/3-540-69710-1_9
32. Young, A., Yung, M.: Bandwidth-optimal kleptographic attacks. In: Koç, Çetin Kaya., Naccache, D., Paar, C. (eds.) Cryptographic Hardware and Embedded Systems – CHES 2001. Lecture Notes in Computer Science, vol. 2162, pp. 235–250. Springer Berlin Heidelberg, Germany, Paris, France (May 14–16, 2001). https://doi.org/10.1007/3-540-44709-1_20
33. Young, A., Yung, M.: Kleptography from standard assumptions and applications. In: Garay, J.A., De Prisco, R. (eds.) SCN 10: 7th International Conference on Security in Communication Networks. Lecture Notes in Computer Science, vol. 6280, pp. 271–290. Springer Berlin Heidelberg, Germany, Amalfi, Italy (Sep 13–15, 2010). https://doi.org/10.1007/978-3-642-15317-4_18
34. Young, A.L., Yung, M.: Backdoor attacks on black-box ciphers exploiting low-entropy plaintexts. In: Safavi-Naini, R., Seberry, J. (eds.) ACISP 03: 8th Australasian Conference on Information Security and Privacy. Lecture Notes in Computer Science, vol. 2727, pp. 297–311. Springer Berlin Heidelberg, Germany, Wollongong, NSW, Australia (Jul 9–11, 2003). https://doi.org/10.1007/3-540-45067-X_26
35. Young, A.L., Yung, M.M.: Space-efficient kleptography without random oracles. In: Furon, T., Cayre, F., Doërr, G., Bas, P. (eds.) Information Hiding. pp. 112–129. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
36. Zhao, Y., Liang, K., Zhao, Y., Yang, B., Ming, Y., Panaousis, E.: Practical algorithm substitution attack on extractable signatures. Designs, Codes and Cryptography **90**(4), 921–937 (2022). https://doi.org/10.1007/s10623-022-01019-1