On the Security of Symmetric Encryption Against Mass Surveillance

DA-ZHI SUN\textsuperscript{1} AND YI MU\textsuperscript{2}, (Senior Member, IEEE)
\textsuperscript{1}Tianjin Key Laboratory of Advanced Networking (TANK), College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
\textsuperscript{2}Fujian Provincial Key Laboratory of Network Security and Cryptology, College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350007, China

Corresponding author: Da-Zhi Sun (sundazhi@tju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872264.

ABSTRACT
For mass surveillance, the algorithm substitution attacks (ASAs) are serious security threats to the symmetric encryption schemes. At CRYPTO 2014, Bellare, Paterson, and Rogaway (BPR) formally developed the security notions of decryptability, undetectability, and surveillance and presented a unique ciphertext symmetric encryption scheme against all possible ASAs. At FSE 2015, Degabriele, Farshim, and Poettering (DFP) relaxed the correctness of decryptability and presented an input-triggered ASA, which meets the BPR security definitions but violates the security of the BPR unique ciphertext scheme. Hence, DFP refined the security notions of detectability and subversion resistance to remove their ASA from the BPR unique ciphertext scheme. At CCS 2015, Bellare, Jaeger, and Kane (BJK) also developed the security notion of key recovery to make the input-triggered ASA infeasible. We investigate ASAs on the symmetric encryption scheme. Our contribution is twofold. (1) We propose a new trigger ASA against the symmetric encryption scheme. Our proposed ASA cannot be captured by the BJK security definitions. Comparatively, the DFP security definitions can detect our proposed ASA. In the view of ASAs, this result demonstrates that the DFP security definitions are not identical to the BJK security definitions. (2) We improve the DFP definition of subversion resistance. DFP proved that the BPR unique ciphertext scheme defeats the input-triggered ASA under their subversion resistance definition. However, we show that the BPR unique ciphertext scheme fails to meet the DFP subversion resistance definition due to our proposed ASA. Therefore, an improved definition on subversion resistance is proposed to cover all existing trigger ASAs. We prove that the BPR unique ciphertext scheme is secure under our improved definition. Therefore, we believe that our improved definition is more suitable to evaluate the ASA security of the symmetric encryption scheme.

INDEX TERMS
Mass surveillance, algorithm substitution attack, symmetric encryption, decryptability, undetectability/detectability, key recovery, subversion resistance.

I. INTRODUCTION
A growing number of cryptographic components are provided by the third party suppliers. One example is that the TLS record layer integrated in Microsoft’s Internet Explorer and Apple’s Safari browsers employs the AES-CBC encryption scheme. Another example is that IPsec using the triple DES-CBC encryption scheme supports Cisco IOS-based routers or Huawei AR G3 routers to protect the confidentiality of IP packets sent over a network. In these examples, the malicious adversary has many opportunities to substitute those encryption schemes in the implementation level.

As a scenario, the closed-source software employs a standard symmetric encryption scheme to provide the confidential protection for the data. For the algorithm substitution attack (ASA), the adversary replaces the executable code of the standard scheme with the executable code of his alternative scheme. The adversary and his alternative scheme are respectively called as the big brother (BB) and the subversion. A successful ASA can undermine the confidentiality of the data processed by the standard scheme and at the same time circumvent detection by its honest users. Figure 1 gives a brief overview of this procedure.

In [1], Bellare, Paterson, and Rogaway (BPR) explored ASAs on several well-known symmetric encryption schemes and formalized the security model of ASAs. They further
designed a unique ciphertext symmetric encryption scheme, which proves to be secure under the BPR security model. Later, Degabriele, Farshim, and Poettering (DFP) [2] showed that a weakened BPR security definition renders possible an input-triggered ASA on the BPR unique ciphertext scheme. They refined the BPR security model to restore the positive result of the BPR unique ciphertext scheme. In 2015, Bellare, Jaeger, and Kane (BJK) [3] presented a stateless ASA, which can break all randomized symmetric encryption schemes. They also enhanced the BPR security model [1] and showed that their security model can invalidate DFP’s input-triggered ASA.

A. CONTRIBUTION
This paper aims to further investigate ASAs, its security models, and its schemes under the symmetric encryption setting. We reveal a new and efficient trigger ASA and improve the DFP security definitions owing to our proposed ASA and BPR unique ciphertext scheme. Our contribution is further summarized as follows.

(1) We propose a new ASA, which uses a public parameter as the trigger. Unlike the DFP input-triggered ASA, our proposed ASA does not require BB to input any trigger data into the encryption algorithm. On the one hand, we show that the BJK security model fails to capture our proposed ASA. That is, our proposed ASA satisfies the BJK security definitions under their strong undetectability game and key recovery game. On the other hand, the DFP security definitions under their surveillance game and detection game can detect our proposed ASA. Hence, the DFP security definitions and the BJK security definitions are not identical with respect to our proposed ASA.

(2) We improve the DFP security definition to resist ASAs. DFP defined the subversion resistance to prevent all ASAs. They proved that the BPR unique ciphertext scheme meets their security definition and is thus secure. However, we demonstrate that the scheme is instead not secure as to our proposed ASA. We therefore improve their definition. The BPR unique ciphertext scheme is secure again under our improved definition. It means that our improved definition is more suitable to evaluate the ASA security of the symmetric encryption scheme, compared with the DFP security definition.

B. RELATED WORK
In [4], [5], the subliminal channels enlightened by the prisoners’ problem are regarded as an early form of ASAs. Young and Yung [6]–[8] extended the subliminal channel threat to a broader framework of kleptography. Backdoored blockciphers were examined in [9], [10]. Goh et al. [11] showed how to add key recovery to existing security protocols such as SSL/TLS and SSH without changing the protocol. Waksman and Sethumadhavan [12] dedicated to preventing the hidden backdoors in hardware components.

Snowden revealed numerous global surveillance programs [13], [14] run by the NSA (National Security Agency) and the Five Eyes Intelligence Alliance. In fact, his disclosures refueled the research over this area. Some literature [1]–[3], [15], [16] focused on studying the ASA problem on symmetric encryption schemes. Russell et al. [17] generalized ASAs by permitting adversarial subversion of (randomized) key generation. Dodis et al. [18] presented a formal treatment of backdoored pseudorandom generators (PRGs), which can be treated as a subversion. Mironov and Stephens-Davidowitz [19] presented cryptographic reverse firewalls to counter ASAs via trusted code in network perimeter. The cryptographic reverse firewall is a generic way to prevent a tampered machine from leaking information to BB via any scheme. And the cryptographic reverse firewalls [20]–[22] mainly targeted at the protection of the public-key schemes.
According to the results in [20], the improved asymmetric subversion model for signature and identification [23], [24] is further presented. The countermeasures against hardware trojans were proposed by Dziembowski et al. [25] and Atieniese et al. [26]. The hardware trojan can also be treated as an example of ASAs. Bellare et al. [27] suggested to thwart the symmetric secret key exfiltration by an enormously long key. This idea also helps to defeat ASAs. Fischlin and Mazaberi [28] put forward the self-guarding constructions for basic cryptographic primitives against ASAs. Berndt and Liśkiewicz [29] proved that successful ASAs correspond to secure stegosystems on certain channels and vice versa. Giacon et al. [30] proposed key-encapsulation mechanism (KEM) combiners, which can be potentially employed to prevent ASAs. Auerbach et al. [31] studied the security of public-key encryption schemes and KEMs when public parameters they use may be subverted. Armour and Poettering [32] studied options to subvert symmetric encryption algorithm, and the external accomplice mechanism (KEM) combiners, which can be potentially subverted. Due to multiple surveillants for different reasons, Giacon et al. [33] showed an efficient way to undetectably subvert the well-known lattice-based encryption scheme proposed by Regev. Baek et al. [34] presented a highly efficient ASA on the digital signature algorithm (DSA) and implemented the proposed ASA by replacing the original DSA in Libgcrypt with the subverted DSA. Due to multiple surveillants for different governments or manufacturers, Li et al. [35] initializes the analysis of security against subversion in a multi-surveillant setting. They introduced a security notion that the transmission of a real message is undetectable, which means all surveillants either think the users transmit an innocuous message by the subverted algorithms, or consider users are using non-subverted algorithms. Lv et al. [36] investigated the study of subversion attacks against cloud auditing protocol. In addition, Schneier et al. [37] categorized a broader set of potential avenues for the subversion of cryptographic systems.

II. SYMMETRIC ENCRYPTION SCHEME AND ITS ASA

This section uses the formal method to describe the symmetric encryption scheme and its ASA. Next, two crucial concepts, i.e., the correctness and the decryptability, are respectively defined in the formal sense. The correctness denotes the error rate of the encryption and decryption process without considering the ASA. Comparatively, the decryptability denotes the error rate of the encryption and decryption process under the ASA.

A. SYMBOLS AND NOTATIONS

We follow the symbols and notations in [1], [2] to discuss the symmetric encryption scheme and its subversion.

- \( \{0, 1\}^n \) denotes the set of \( n \)-bit strings.
- \( \{0, 1\}^* \) denotes the set of all strings.
- \( \perp \) denotes a special symbol standing for “invalid” or “reject”.
- \( \varepsilon \) denotes the empty string.
- \( \mathbb{N} \) denotes the set of the natural numbers.

- \( \Pi = (K, E, D) \) and \( \tilde{\Pi} = (\tilde{K}, \tilde{E}, \tilde{D}) \) denote the symmetric encryption scheme and its subversion. \( K, E, \) and \( D \) in \( \Pi \) are respectively the key space, the encryption algorithm, and the decryption algorithm. \( \tilde{K}, \tilde{E}, \) and \( \tilde{D} \) in \( \tilde{\Pi} \) are respectively the subversion key space, the subversion encryption algorithm, and the external accomplice algorithm.
- \( M, C, \) and \( AD \) respectively denote the message space, the ciphertext space, and the associated data space.
- \( |S| \) represents the size of the finite set \( S \).
- \( S \not\in S \) represents picking \( S \) at random from a set \( S \).
- \( y \leftarrow A(x_1, x_2, \ldots) \) represents running an algorithm \( A \) (perhaps with random coins) on inputs \( x_1, x_2, \ldots \) to deterministically obtain output \( y \).
- \( \Pr[X] \) represents the probability of an event \( X \).
- \( \Pr[P : X] \) represents the probability of an event \( X \) occurring after having executed a process \( P \).
- \( \Pr[X|Y] \) represents the conditional probability of an event \( X \) occurring given that an event \( Y \) has occurred.

B. SYMMETRIC ENCRYPTION SCHEME

A symmetric encryption scheme is a triple \( \Pi = (K, E, D) \). \( K \) is a non-empty set of strings. \( K \) has a message space \( M \subseteq \{0, 1\}^* \) and a ciphertext space \( C \subseteq \{0, 1\}^* \). To accommodate the authenticated encryption, \( \Pi \) has an associated data space \( AD \subseteq \{0, 1\}^* \). \( E \) may be randomized. To be more specific, \( E \) maps the secret key \( K \in K \), a message \( M \in M \), associated data \( A \in AD \), and the current state \( \sigma \) to a ciphertext \( C \in C \) (or \( \perp \)) and the updated state \( \sigma' \). That is, \( (C, \sigma') \leftarrow E_K(M, A, \sigma, D) \) is a deterministic algorithm that maps the secret key \( K \in K \), a ciphertext \( C \in C \), associated data \( A \in AD \), and the current state \( \sigma \) to a message \( M \in M \) and the updated state \( \sigma' \). That is, \( (M, \sigma') \leftarrow D_K(C, A, \sigma) \). The encryption and decryption states are always initialized to \( \varepsilon \). To understand the state bonded with the symmetric encryption scheme, the counter-mode with zero-initialized counter [1] is a typical example and the counter value is its state.

In an unambiguous context, we may respectively write \( E_K(M, A, \sigma) \) and \( D_K(C, A, \sigma) \) to represent \( C \) and \( M \). For any \( \Pi = (K, E, D) \) and any \( l \in \mathbb{N} \), \( M = \{M_1, \ldots, M_l\} \) and \( A = [A_1, \ldots, A_l] \) are vectors written in bold font, where \( M_1, \ldots, M_l \in M \) and \( A_1, \ldots, A_l \in AD \). Then, \( (C_1, \sigma_1) \leftarrow E_K(M, A, \varepsilon) \) represents

\[
(C_1, \sigma_1) \leftarrow E_K(M_1, A_1, \varepsilon); \ldots; (C_l, \sigma_l) \\
\leftarrow E_K(M_l, A_l, \sigma_{l-1}),
\]

where \( C = [C_1, \ldots, C_l] \) and \( C_1, \ldots, C_l \in C \). And, \( (M_1', \sigma_1') \leftarrow D_K(C, A, \varepsilon) \) denotes the similar process for decryption, that is,

\[
(M_1', \sigma_1') \leftarrow D_K(C_1, A_1, \varepsilon); \ldots; (M_l', \sigma_l') \\
\leftarrow D_K(C_l, A_l, \sigma_{l-1}'),
\]

where \( M' = [M_1', \ldots, M_l'] \).
We review the notion of correctness for the symmetric encryption scheme as follows.

**Definition 1 (Definition 1 in [2]):** Let \( l, q \in \mathbb{N} \). A symmetric encryption scheme \( \Pi = (\mathcal{K}, \mathcal{E}, \mathcal{D}) \) is said to be \((q, \delta_1)\)-correct if for all \( l \leq q \), any vector \( \mathbf{M} = [M_1, \ldots, M_l] \), and any vector \( \mathbf{A} = [A_1, \ldots, A_l] \), it holds that:
\[
\text{Pr}[K \xleftarrow{\$} \mathcal{K}; (C, \sigma) \leftarrow \mathcal{E}(K, \mathbf{A}, \varepsilon); (M', \sigma') \leftarrow \mathcal{D}(K, \mathbf{A}, \varepsilon) : \mathbf{M} \neq \mathbf{M'}] \leq \delta_1.
\]

C. ASA

A subversion of the scheme \( \Pi = (\mathcal{K}, \mathcal{E}, \mathcal{D}) \) is a triple \( \tilde{\Pi} = (\tilde{\mathcal{K}}, \tilde{\mathcal{E}}, \tilde{\mathcal{D}}) \). \( \tilde{\mathcal{K}} \) is a non-empty set of strings. BB selects the subversion key \( \tilde{K} \in \tilde{\mathcal{K}} \) to hide its malicious behavior. \( \tilde{\mathcal{E}} \) maps the encryption key \( K \in \mathcal{K} \), the subversion key \( \tilde{K} \in \tilde{\mathcal{K}} \), a message \( M \in \mathcal{M} \), associated data \( A \in \mathcal{AD} \), and the current state \( \sigma \) to a ciphertext \( C \in \mathcal{C} \) and the updated state \( \sigma' \). That is, \((C, \sigma') \leftarrow \tilde{\mathcal{E}}[K, \tilde{K}, M, A, \sigma] \). \( \tilde{\mathcal{E}} \) may call \( \mathcal{E} \) in a certain way. \( \tilde{\mathcal{E}} \) receives the subverted \( C \) generated by \( \tilde{\mathcal{E}} \) and aims to violate the security of \( \Pi \). For example, \( \tilde{\mathcal{E}} \) may try to recover the messages or the encryption key from the subverted ciphertexts. Here, \( \tilde{\mathcal{E}} \) and \( \tilde{\mathcal{D}} \) share \( \tilde{K} \). To implement an ASA, BB covertly replaces the executable code of \( \mathcal{E} \) with that of \( \tilde{\mathcal{E}} \). Then, BB listens to the subverted \( C \) over public channel and compromises the security of \( \Pi \).

We do not follow the plaintext-recovery algorithm \( \tilde{\mathcal{D}} \) defined in [1], [2] to describe the severity of the ASA. Instead, \( \tilde{\mathcal{E}} \) is used as in [3] but only to emphasize that BB does not require the substitution of the executable code for the decryption algorithm \( \mathcal{D} \). Assume that \((C, \tilde{\sigma}) \leftarrow \tilde{\mathcal{E}}[K, \tilde{K}, M, A, \sigma] \) denotes the similar encryption process as \((C, \sigma) \leftarrow \mathcal{E}(K, M, A, \sigma) \). For completeness, we restate the notion of decryptability for the subversion encryption algorithm.

**Definition 2 (Definition 3 in [2]):** Let \( l, q \in \mathbb{N} \). We say that a subversion \( \tilde{\Pi} = (\tilde{\mathcal{K}}, \tilde{\mathcal{E}}, \tilde{\mathcal{D}}) \) satisfies \((q, \delta_2)\)-decryptability with respect to a symmetric encryption scheme \( \Pi = (\mathcal{K}, \mathcal{E}, \mathcal{D}) \) if for all \( l \leq q \), any vector \( \mathbf{M} = [M_1, \ldots, M_l] \), and any vector \( \mathbf{A} = [A_1, \ldots, A_l] \), it holds that:
\[
\text{Pr}[K \leftarrow \mathcal{K}; \tilde{K} \leftarrow \tilde{\mathcal{K}}; (C, \tilde{\sigma}) \leftarrow \tilde{\mathcal{E}}[K, \tilde{K}, M, A, \sigma]; (M', \sigma') \leftarrow \tilde{\mathcal{D}}(K, \mathbf{A}, \varepsilon) : \mathbf{M} \neq \mathbf{M'}] \leq \delta_2.
\]

Practically, we distinguish two possible cases in terms of Definitions 1 and 2.

1. \( \delta_1 = 0 \) and \( \delta_2 = 0 \) for all \( q \in \mathbb{N} \). In [1]–[3], both \( \Pi \) and its decryptability with \( \tilde{\Pi} \) are said to be perfectly correct in this case.

2. \( \delta_2 \neq 0 \) but \( \delta_2 \) is a negligible value for any reasonable \( q \in \mathbb{N} \). BPR required that any \( \Pi \) satisfy the perfect decryptability, i.e., \( \delta_2 = 0 \), and designed a unique ciphertext scheme \( \Pi \) against all existing \( \tilde{\Pi} \). However, DFP argued against the perfect decryptability condition and proposed an input-triggered \( \Pi \) on the BPR unique ciphertext scheme \( \Pi \) when \( \delta_2 \) is a negligible value. DFP proposed a security model to capture their input-triggered \( \Pi \). They further showed that the BPR unique ciphertext scheme \( \Pi \) is secure under the DFP security model. BJK [3] also did not require the perfect decryptability condition and argued that:

   *We have dropped this condition, so that decryptability holds only to the extent that it is implied by strong undetectability, which we think is more realistic from a detection perspective.*

III. ASA SECURITY MODELS

To evaluate ASAs, the formal security models are always constructed under the interaction games, which relate to the subversions. The security requirements for ASAs are defined by using these games. Here, we review the BJK security model [3] and the DFP security model [2].

A. THE BJK SECURITY MODEL

The BJK security model focuses on the security definitions of the acceptable subversions. Therefore, a symmetric encryption scheme is secure against ASAs, if there does not exist the subversion on the scheme where the subversion satisfies the BJK security definitions. In the BJK security model, associated data \( A \) is omitted. We follow this style, when the BJK security model is used.

1) STRONG UNDETECTABILITY

As shown in the left column of Fig. 2, the strong undetectability is formalized by the game SDET. The SDET is associated with a benign detection adversary \( SD \). A random bit \( b \) and subversion key \( \tilde{K} \) are first sampled. \( SD \) then has access to the encryption oracle \( Enc \). Upon receiving \((K, M, A)\), the oracle \( Enc \) produces \((C, \sigma)\) either via \( \mathcal{E} (b = 1) \) or via \( \tilde{\mathcal{E}} (b = 0) \). \( SD \) needs to determine \( b \). The detection advantage of \( SD \) is defined as

\[
\text{Adv}^{\text{sdet}}_{\Pi, \tilde{\Pi}}(SD) = 2\text{Pr}[\text{SDET}^{SD}_{\Pi, \tilde{\Pi}} = \text{true}] - 1.
\]

The BJK security model mainly considers the stateless \( E \), that is, \( \sigma \) returned by \( E \) is always the empty string \( \varepsilon \). Thus, \( \tilde{E} \) must keep stateless to achieve the strong undetectability. Otherwise, \( SD \) can determine that \( b = 0 \), if \( \sigma \neq \varepsilon \) in a reply to the Enc query.

2) KEY RECOVERY

The game KR in the right column of Fig. 2 is used to evaluate the effectiveness of key recovery. \( \tilde{\Pi} \) wins if \( \tilde{E} \) recovers the...
encryption key $K$ from the ciphertexts $C$ produced by $\tilde{E}$. The key recovery advantage of $\tilde{L}$ is defined as

$$\text{Adv}_{\tilde{L}}^{\text{kr}}(\tilde{E}) = \Pr[KR_{\tilde{L},\tilde{M}S} = \text{true}].$$

Here, a message sampler algorithm $\tilde{M}S$ represents the choice of messages made by the honest user. That is, its current state $\sigma''$, $\tilde{M}S$ returns the next message $M$ to be encrypted and updates its $\sigma'$. For key recovery attacks, the less they assume about $\tilde{M}S$, the stronger they are. The strongest attack should work for any $\tilde{M}S$.

An acceptable subversion $\tilde{\Pi}$ requires that the strong undetectability advantage $\text{Adv}_{\tilde{\Pi},\tilde{\Pi}}^{\text{sdet}}(SD)$ be negligible but the key recovery advantage $\text{Adv}_{\tilde{\Pi},\tilde{M}S}^{\text{kr}}(\tilde{L})$ not.

### B. THE DFP SECURITY MODEL

The DFP security model is taking particular aim at the security definitions of detecting the subversions on symmetric encryption schemes. That is, any subversion should be efficiently identified by the benign test, once BB executes the ASA on the symmetric encryption scheme. Therefore, the subversions should fail, because BB is afraid of being detected.

#### 1) SURVEILLANCE

As shown in the right column of Fig. 3, BB associated with the surveillance game $\text{SURV}$ is called as $B$. The surveillance game $\text{SURV}$ starts by randomly generating a bit $b$ and subversion key $\tilde{K}$. Given $\tilde{K}$, $B$ has access to the key generation oracle Key and the encryption oracle $\text{Enc}$. Depending on the value of $b$, the oracle $\text{Enc}$ returns the ciphertexts $C$ generated by either $\tilde{E}$ ($b = 1$) or $\tilde{E}$ ($b = 0$). $B$ outputs a bit $b'$ as the guess of the challenge $b$. The surveillance advantage of $B$ is given by

$$\text{Adv}_{\Pi,\tilde{\Pi}}^{\text{sur}}(B) = 2\Pr[\text{SURV}_{\Pi,\tilde{\Pi}} = \text{true}] - 1.$$

#### 2) DETECTABILITY

The detection game $\text{DETECT}$ is an extension of the surveillance game $\text{SURV}$. First, given with the subversion key $\tilde{K}$, $B$ runs the surveillance game $\text{SURV}$ to subvert the encryption key $K$. Simultaneously, its encryption queries are recorded in a transcript $T$. Then, the detection test $U$ is only given access to $T$, which includes $K$. The goal of $U$ is to output a bit $b''$ as the guess of the random bit $b$. See the left column of Fig. 3 for the game $\text{DETECT}$. The detection advantage of $U$ is given by

$$\text{Adv}_{\Pi,\tilde{\Pi}}^{\text{det}}(B, U) = 2\Pr[\text{DETECT}_{\Pi,\tilde{\Pi}} = \text{true}] - 1.$$

Clearly, DFP specified the security notion of the detectability rather than the strong undetectability. Moreover, in the BJK security model and the DFP security model, both $SD$ and $U$ have access to the encryption oracle $\text{Enc}$ in a certain way. Of course, these two security models do not capture all detection methods, for example, the side-channel attacks [38], [39].

### IV. PROPOSED ASA

#### A. SUBVERSION DESCRIPTION

We are now to construct a subversion $\Pi$ for any scheme $\Pi$. For simplicity, we assume that the message and key are of the same size. (However, the subversion can be easily extended to cover the case, where the key has a larger size.) $\tilde{E}$ and $\tilde{L}$ of $\Pi$ are shown in Fig. 4. Our $\Pi$ introduces a variable $\text{pubinform}$, which is always available to all components of the cryptographic system. Let $\text{pubinform}$ be independent of all cryptographic operations, i.e., encryption, decryption, and subversion encryption. The function $\text{Setcurrentpubinform}(\text{val})$ sets $\text{pubinform}$ to $\text{val}$. The function $\text{Getcurrentpubinform}()$ returns the current value of $\text{pubinform}$. Assume that $\text{pubinform}$ has variable space $\{0, 1\}^{\lambda}$. BB selects a value $\text{PUBINFORM} \in \{0, 1\}^{\lambda}$, where $\Pr(\text{pubinform} = \text{PUBINFORM} \text{ when no manipulation for pubinform})$ is negligible. And, $\text{PUBINFORM}$ is secretly shared by $\tilde{E}$ and $\tilde{L}$.
In our $\overline{\Pi}$, PUBINFORM is used as the trigger for leaking the encryption key $K$ of $\Pi$ to BB. We describe $\tilde{E}$ and $\tilde{E}$ in detail.

$\tilde{E}$, $K$, $\tilde{K}$, $M$, $A$, and $\sigma$ are inputs. The encryption process of $\tilde{E}$ is as follows. $\tilde{E}$ with PUBINFORM first gets the current value of pubinform by calling Getcurrentpubinform(). If $\tilde{E}$ = PUBINFORM, $\tilde{E}$ computes $\tilde{E}_K(K, A, \sigma)$ as its output; otherwise $\tilde{E}$ behaves exactly like the desired $E$ in $\Pi$.

$\tilde{L}$, $\tilde{K}$ and PUBINFORM are inputs. $\tilde{L}$ sets the current value of pubinform by calling Setcurrentpubinform(PUBINFORM). $\tilde{L}$ calls $\tilde{E}$, when pubinform = PUBINFORM. Here, $\tilde{E}(\tilde{K}, M, A, \sigma)$ means that BB does not need any input for key recovery. Then, $\tilde{E}$ returns $C$ to $\tilde{L}$, where $C = \tilde{E}_K(K, A, \sigma)$. $\tilde{L}$ eventually obtains $K$ by computing $D_{\tilde{K}}(C, A, \sigma)$ (= $D_{\tilde{K}}(\tilde{E}_K(K, A, \sigma), A, \sigma) = K$).

Remark 1 (About Trigger pubinform): Although the encryption module can be regarded as a black box to BB during the system running, BB can tamper with other unprotected public parameters of the system for the subverting purpose. For a cryptographic chip, the system clock can be taken as the trigger. In general, the system clock can be configured. Assume that the system clock uses the 16-bit pubinform to represent the current year between AD 1 and AD 65535. That is, pubinform = 0000000000000001 denotes AD 1, pubinform = 00000111111100001 denotes AD 2018, and pubinform = 00000111111110011 denotes AD 2019, etc. BB may choose PUBINFORM = 11111111111111111. This PUBINFORM means AD 65535, the largest year that pubinform can represent. To recover the encryption key $K$ of $\Pi$, BB can set his chosen PUBINFORM to pubinform and call our proposed $\tilde{E}$. Another example is that some secure system employs 20-byte of ASCII code, i.e., pubinform, to denote the user name. As the user name is always made up of printable characters, BB can choose the non-printable characters in ASCII code as his PUBINFORM. In these examples, each pubinform is independent of the encryption, decryption, subversion encryption operations. At the same time, each Pr(pubinform = PUBINFORM when no manipulation for pubinform) can be treated as a negligible value. Clearly, BB is able to recover the encryption key $K$ using $\tilde{E}$. In practice, a sophisticated BB may reset pubinform to its original value after his key recovery action. It is, however, not easy for the detector without observing pubinform to figure out the existence of $\tilde{E}$.

Remark 2 (State of $\tilde{E}$): In the view of BJK [3], the stateful $\tilde{E}$ should maintain the subversion state $\sigma$ across invocations. This state $\sigma$ is perhaps an integer either representing which bit of the encryption key $K$ $\tilde{E}$ is trying to exfiltrate, or taking a special value to indicate that exfiltration is complete and encryption process should be done as usual. Our proposed $\tilde{E}$ checks the current value of pubinform using Getcurrentpubinform() to determine whether $K$ needs to be exfiltrated. However, $\tilde{E}$ does not need to maintain pubinform. Hence, our proposed $\tilde{E}$ is stateless, that is, the output $\sigma$ of $\tilde{E}$ in the game SDET (see Fig. 2) should always be $\varepsilon$.

Remark 3 (Discussion on Our and DFP’s ASAs): To the best of our knowledge, DFP’s ASA [2] is the only trigger ASA on symmetric encryption scheme. Other known ASAs on symmetric encryption scheme can be repaired by the BPR unique ciphertext scheme [1]. Hence, we merely focus on our and DFP’s ASA. In DFP’s ASA, BB randomly chooses a message $M$ as the subversion key, i.e., $\tilde{K} = M$. A predicate $R_{\tilde{K}, \tilde{K}}(M, A, \sigma)$ used in $\tilde{E}$ takes a true value if $M = \tilde{K}$ and a false value otherwise. Once $\tilde{E}$ receives the message $M = \tilde{K}$ and checks $R_{\tilde{K}, \tilde{K}}(\tilde{K}, A, \sigma) = true$, the encryption key $K$ as a part of the output is returned to $\tilde{E}$. Hence, BB must manipulate the input of $\tilde{E}$ to obtain $K$. DFP further stated that:

"In case the trigger is the set of inputs for which the predicate $R$ holds. In practice, $R$ can depend on any information that the subverted encryption algorithm may have access to, such as an IP address, a username, or some location information. Such information, in particular network addresses and routing information, can be readily available in the associated data. It is not unreasonable, and is in fact in accordance with the usual approach adopted in cryptography, to assume that big brother may be capable of influencing this information when it needs to intercept a communication."

The above argument implied that the associated data $A$ also can be exploited as the trigger. However, if the inputs of $\tilde{E}$ cannot be controlled by BB, such input-triggered ASAs may fail. In fact, BJK [3] used this idea to thwart the input-triggered ASAs. That is, the game KR in Fig. 2 does not allow the encryption oracle Enc to receive any input from BB but uses the message $M$ generated by the sampler algorithm $MS$. In this situation, the probability of the event $\tilde{K}$ is negligible. Comparatively, our proposed ASA is not an input-triggered ASA. It is clear that, when BB employs $\tilde{E}$ described in Fig. 4 to subvert the encryption key $K$, it does not need to input the message $M$ or associated data $A$ into $\tilde{E}$ anymore. Our proposed ASA instead exploits the publicly available pubinform as the trigger. After setting the current value of pubinform, BB calls $\tilde{E}$ to recover the encryption key $K$.

Finally, the performance comparisons between DFP’s ASA and our proposed ASA are summarized in Table 1.

**B. SECURITY PROPERTIES**

Since our proposed ASA is a new trigger ASA on the symmetric encryption schemes, we need to understanding the
theoretical results of our proposed ASA. In the following, we therefore analyze our proposed ASA under the BJK security model and the DFP security model, respectively. Moreover, our proposed ASA can be used as the acid test of the validity of both the BJK security model and the DFP security model.

We firstly recall the security notion of the standard privacy [40]–[42]. For \( \Pi = (K, \mathcal{E}, \mathcal{D}) \), consider the game PRIV depicted in Fig. 5. The advantage of the adversary \( A \) is defined as

\[
\text{Adv}_{\Pi}^{\text{PRIV}}(A) = 2\Pr[\text{PRIV}_{\Pi}^A = \text{true}] - 1.
\]  

(2)

\( \Pi \) is said to be \( \nu \)-private, if for every practical \( A \) its advantage \( \text{Adv}_{\Pi}^{\text{PRIV}}(A) \) is bounded from above by a negligible value \( \nu \). We require \( \Pi \) to be secure in this sense, so that we can restrain our attention to subversion security in the following discussion.

---

**TABLE 1. Performance comparison: DFP’s ASA vs. Ours.**

| ASA              | Access to input of encryption operation | Access to unprotected public parameter of system | Efficiency of key recovery | State of subversion | Difficulty of attack implementation |
|------------------|----------------------------------------|-----------------------------------------------|--------------------------|-------------------|-----------------------------------|
| DFP’s ASA        | Yes                                     | No                                            | Efficient                | Stateless          | Slightly difficult                |
| Our proposed ASA | No                                      | Yes                                           | Efficient                | Stateless          | Easy                              |

---

\[ \tilde{\mathcal{E}} \] makes the corresponding cryptographic operations for the message vector \( \mathbf{M} \) as in Definition 2. We know \( \text{pubinform} = \text{PUBINFORM} \) and the cryptographic operations are independent events. Let \( \tilde{E}_1 \) denote the complementary event of \( E_1 \), that is, \( \text{pubinform} \neq \text{PUBINFORM} \) when \( \tilde{E} \) makes the cryptographic operations for \( \mathbf{M} \). According to Definition 2, we estimate the error probability of decryptability as follows.

\[
\delta_2 = \Pr[K \leftarrow \tilde{K}; (\mathcal{C}, \varepsilon) \leftarrow \tilde{\mathcal{E}}_{\tilde{K}}(\mathbf{M}, \varepsilon);
(\mathcal{M}', \varepsilon) \leftarrow \mathcal{D}_\mathcal{K}(\mathcal{C}, \varepsilon) : \mathbf{M} \neq \mathbf{M}'] = \Pr[\mathbf{M} \neq \mathbf{M}'|\tilde{E}_1]
\]

\[ + \Pr[\mathbf{M} \neq \mathbf{M}'|E_1] \leq \Pr[\mathbf{M} \neq \mathbf{M}'|\tilde{E}_1] + \Pr[E_1] \leq \delta_1 + 1 - (1 - \delta_3)^q \leq \delta_1 + q\delta_3, \]

where the bound of the term \( \Pr[\mathbf{M} \neq \mathbf{M}'|\tilde{E}_1] \) follows from the correctness probability of \( \Pi \).

Consider that the game SDET in Fig. 2 employs \( \mathcal{E} \) and \( \tilde{\mathcal{E}} \), where \( \tilde{\mathcal{E}} \) is described in Fig. 4. \( \mathcal{S} \mathcal{D} \) plays with the encryption oracle \( \text{Enc} \) to guess the random bit \( b \). If \( b = 1 \), \( \mathcal{S} \mathcal{D} \) interacts with \( \mathcal{E} \); otherwise \( \mathcal{S} \mathcal{D} \) runs with \( \tilde{\mathcal{E}} \). Note that \( \tilde{\mathcal{E}} \) calls \( \mathcal{E} \) except in the case when \( \text{pubinform} = \text{PUBINFORM} \). We know that \( \mathcal{S} \mathcal{D} \) in the game SDET never calls \( \text{Setcurrentpubinform}(\text{PUBINFORM}) \) to set the value of \( \text{pubinform} \). For any \( \mathcal{S} \mathcal{D} \) making at most \( q_1 \) encryption queries, it is easy to see that the detection advantage of \( \mathcal{S} \mathcal{D} \) is

\[
\text{Adv}_{\Pi,\mathcal{M}^{(6)}}(\mathcal{S} \mathcal{D}) \leq 1 - (1 - \delta_3)^{q_1} \leq 1 - (1 - q_1\delta_3) = q_1\delta_3.
\]

Consider the game KR in Fig. 2, which uses \( \tilde{\mathcal{L}} \) described in Fig. 4 to recover the encryption key \( K \). We know that \( \tilde{\mathcal{L}} \) can always set the value of \( \text{pubinform} \) by using \( \text{Setcurrentpubinform}(\text{PUBINFORM}) \) and further query the encryption oracle \( \text{Enc} \). Moreover, \( \text{PUBINFORM} \) is not an input of \( \tilde{\mathcal{E}} \). Hence, \( \tilde{\mathcal{E}} \) should compute \( \tilde{\mathcal{E}}_{\tilde{K}}(K, \varepsilon) \) and return the corresponding \( \mathcal{C} = \tilde{\mathcal{E}}_{\tilde{K}}(K, \varepsilon) \) to \( \tilde{\mathcal{L}} \). It means that \( \tilde{\mathcal{L}} \) can recover \( K \) by computing \( \mathcal{D}_{\mathcal{K}}(\mathcal{C}, \varepsilon) = \mathcal{D}_{\mathcal{K}}(\tilde{\mathcal{E}}_{\mathcal{K}}(K, \varepsilon)) = K \).

Therefore, the key recovery advantage of \( \tilde{\mathcal{L}} \) is 1, i.e.,

\[
\text{Adv}_{\Pi,\mathcal{M}^{(6)}}(\tilde{\mathcal{L}})^{\text{KR}} = 1.
\]

---

**Theorem 1:** Let \( \Pi = (K, \mathcal{E}, \mathcal{D}) \) be a \((1, \delta_1)\)-correct symmetric encryption scheme as in Fig. 5. Let \( \tilde{\Pi} = (\tilde{K}, \tilde{\mathcal{E}}, \tilde{\mathcal{L}}) \) depicted in Fig. 4 be \((q, \delta_2)\)-decryptability relative to \( \Pi \). Assume that the variable \( \text{pubinform} \) is independent of all encryption, decryption, and subversion encryption operations. Let the space of \( \text{pubinform} \) be \( \{0, 1\}^\lambda \) and \( \text{PUBINFORM} \in \{0, 1\}^\lambda \). Assume that \( \Pr[\text{pubinform} = \text{PUBINFORM} \wedge \text{no manipulation for pubinform}] = \delta_3 \), where \( \delta_3 \) is negligible. Then, \( \delta_2 \leq \delta_1 + q\delta_3 \). Under the BJK security model, the detection advantage of \( \mathcal{S} \mathcal{D} \) is

\[
\text{Adv}_{\Pi,\mathcal{M}^{(6)}}(\mathcal{S} \mathcal{D}) \leq q_1\delta_3, \text{ when } \mathcal{S} \mathcal{D} \text{ makes at most } q_1 \text{ encryption queries. Moreover, the key recovery advantage of } \tilde{\mathcal{L}} \text{ is } \text{Adv}_{\Pi,\mathcal{M}^{(6)}}(\tilde{\mathcal{L}})^{\text{KR}} = 1.
\]

**Proof 1:** We omit associated data in this proof. As shown in Fig. 4, \( \tilde{\mathcal{E}} \) should call \( \tilde{\mathcal{E}}_{\mathcal{K}}(M, \varepsilon) \) and return the output of \( \mathcal{E}_{\mathcal{K}}(M, \varepsilon) \) except when the event \( \text{pubinform} = \text{PUBINFORM} \) happens. Consider the case that \( \text{pubinform} \) is not maliciously set by calling \( \text{Setcurrentpubinform}(\text{PUBINFORM}) \). Let \( E_1 \) denote the event \( \text{pubinform} = \text{PUBINFORM} \) happens when...
model, the detection advantage of $\mathcal{U}$ is $\text{Adv}^{\text{det}}_{\text{V. FLAW ON THE DFP SECURITY DEFINITION}}(\mathcal{B}, \mathcal{U}) \geq 1 - v - q_1 \delta_1$, when $\mathcal{B}$ makes at most $q_1$ encryption queries.

Proof 2: Let us consider that the game $\text{DETECT}$ in Fig. 3 employs $\tilde{E}$ and $\tilde{E}$, where $\tilde{E}$ is described in Fig. 4. First, given the subversion key $K$, $\mathcal{B}$ runs with the key generation oracle $\mathcal{K}$ and the encryption oracle $\mathcal{E}_K$, and its corresponding queries are recorded in a transcript $T$. To guess the random bit $b$, $\mathcal{B}$ perhaps uses $\tilde{E}$ described in Fig. 4 to verify the existence of $\tilde{E}$. Then, the detection test $\mathcal{U}$ has access to $T$, which includes the encryption key $K$ and the corresponding encryption triples $\{M, A, C\}$. The goal of $\mathcal{U}$ is to guess the random bit $b$ by $T$.

In Fig. 6, we propose a detection test $\mathcal{U}$ to infer its $b''$ as a guess of $b$. To distinguish $\tilde{E}$ and $\mathcal{E}$, the idea of the detection test $\mathcal{U}$ is to detect the encryption error appeared in the transcript $T$. $\mathcal{U}$ interprets the encryption key $K$ and the encryption triples $\{M, A, C\}$ from $T$. Then, $\mathcal{U}$ decrypts each ciphertext $C$ in $T'$ by computing $\mathcal{D}_K(C, A, \sigma)$. If any message $M$ is not equal to its decrypted value, $\mathcal{U}$ decides that $b = b'' = 0$. That is, $\mathcal{U}$ believes that the encryption oracle $\mathcal{E}$ runs $\tilde{E}$ for the game. Otherwise, $\mathcal{U}$ infers that the oracle $\mathcal{E}$ calls $\mathcal{E}$ and therefore decides that $b = b'' = 1$. We estimate the detection advantage of $\mathcal{U}$. According to Eq. (1), we know the detection advantage of $\mathcal{U}$ is expressed as $\text{Adv}^{\text{det}}_{\text{V. FLAW ON THE DFP SECURITY DEFINITION}}(\mathcal{B}, \mathcal{U}) = 2\text{Pr}[\text{DETECT}^B_{\mathcal{B}, \mathcal{U} \mathcal{I}} = \text{true}] - 1$. The game $\text{DETECT}$ tells us that:

$\text{Pr}[\text{DETECT}^B_{\mathcal{B}, \mathcal{U} \mathcal{I}} = \text{true}] = \text{Pr}[b = b''] = \text{Pr}[b = b' = 0] + \text{Pr}[b = b' = 1] = \text{Pr}[b'' = 0|b = 0]\text{Pr}[b = 0] + \text{Pr}[b'' = 1|b = 1]\text{Pr}[b = 1].$ (3)

Now, we calculate the upper bounds of $\text{Pr}[b'' = 0|b = 0]$ and $\text{Pr}[b'' = 1|b = 1]$.

![Algorithm](image)

When $b = 0$, the encryption oracle $\mathcal{E}_K$ runs $\tilde{E}$ in Fig. 3 and $\mathcal{B}$ should call $\tilde{E}$ to obtain the encryption key $K$. If $\text{pubinform} = \text{PUBINFORM}$, the oracle $\mathcal{E}_K$ should output a ciphertext $C$ which is computed by $\mathcal{E}_K(K, A, \epsilon)$ (not by $\mathcal{E}_k(M, A, \sigma)$). This $\{M, A, C = \mathcal{E}_K(K, A, \epsilon)\}$ is recorded in the transcript $T$. Therefore, if the detection test $\mathcal{U}$ makes a wrong guess, i.e., $b'' = 1$ but $b = 0$, the event $M = \mathcal{D}_K(\tilde{E}_K(K, A, \epsilon), A, \sigma)$ should happen. We argue that

$\text{Pr}[M = \mathcal{D}_K(\tilde{E}_K(K, A, \epsilon), A, \sigma)] \leq v,

where $v$ is the upper bound of the advantage $\text{Adv}^{\text{PRIV}}_{\text{V. FLAW ON THE DFP SECURITY DEFINITION}}(\mathcal{A})$ defined in Eq. (2). Suppose on the contrary that

$\text{Pr}[M = \mathcal{D}_K(\tilde{E}_K(K, A, \epsilon), A, \sigma)] > v.$

The event $M = \mathcal{D}_K(\tilde{E}_K(K, A, \epsilon), A, \sigma)$ tells us that $\mathcal{U}$ without $K$ can correctly decrypt $\tilde{E}_K(K, A, \epsilon)$. Therefore, $\mathcal{U}$ is an adversary $\mathcal{A}$, who can win the game $\text{PRIV}$ depicted in Fig. 5. And, its advantage $\text{Adv}^{\text{PRIV}}_{\mathcal{V. FLAW ON THE DFP SECURITY DEFINITION}}(\mathcal{U})$ is not bounded by $v$. It is a contradiction. Thus, we know

$\text{Pr}[b'' = 0|b = 0] = 1 - \text{Pr}[b'' = 1|b = 0] \geq 1 - \text{Pr}[M = \mathcal{D}_K(\tilde{E}_K(K, A, \epsilon), A, \sigma)] \geq 1 - v.$ (4)

When $b = 1$, $\mathcal{B}$ runs the encryption oracle $\mathcal{E}_K$, which uses $\tilde{E}$ and records $\{M, A, C\}$ in the transcript $T$. If the detection test $\mathcal{U}$ makes a wrong guess, i.e., $b'' = 0$, there is at least an encryption error during the interaction between $\mathcal{B}$ and the oracle $\mathcal{E}_K$. Let $\mathcal{E}_2$ denote the event that there is at least a $\{M, A, C\}$ in $T$ such that $M \neq \mathcal{D}_K(\tilde{E}_K(M, A, \sigma), A, \sigma)$. Based on this observation and Definition 1, we have

$\text{Pr}[b'' = 1|b = 1] = 1 - \text{Pr}[b'' = 0|b = 1] 
\geq (1 - \text{Pr}[\mathcal{K}^{-1}(b'' = 0|b = 1) = (1 - \delta_1)^{q_1} \geq 1 - q_1\delta_1],$ (5)

where $q_1$ is the maximum number of $b$’s encryption queries. Moreover, combining Eq. (1) with Eqs. (3), (4), and (5), we thus get

$\text{Adv}^{\text{det}}_{\text{V. FLAW ON THE DFP SECURITY DEFINITION}}(\mathcal{B}, \mathcal{U}) = 2\text{Pr}[\text{DETECT}^B_{\mathcal{B}, \mathcal{U} \mathcal{I}} = \text{true}] - 1
\geq 2(\text{Pr}[b'' = 0|b = 0]\text{Pr}[b = 0] + \text{Pr}[b'' = 1|b = 1|\text{Pr}[b = 1])
\geq 2(1 - v + 1 - q_1\delta_1) - 1 = 1 - v - q_1\delta_1.$

Here, we know that $\text{Pr}[b = 0] = \text{Pr}[b = 1] = 1/2$, because $b$ is a random bit.

Theorem 2 demonstrates that the game $\text{DETECT}$ in the DFP security model can find out our proposed ASA because $v$ and $q_1\delta_1$ are all negligible. It implies that the BJK security model is incompatible with the DFP security model. That is, in view of detecting ASAs, the BJK security model and the DFP security model are not equivalent.

V. FLAW ON THE DFP SECURITY DEFINITION

In [2], DFP stated their security model can distinguish all ASAs including their input-triggered ASA and proved that the BPR unique ciphertext scheme is secure under their security model. However, we demonstrate that DFP security model has the flaw, because the BPR unique ciphertext scheme is insecure due to our proposed ASA. Hence, we improve DFP security model to address this flaw and use it to restore the positive result of the BPR unique ciphertext scheme. Due to fixing the flaw on the DFP security definition,
our improved security model is more reasonable than DFP security model.

DFP proposed the security notion of subversion resistance against their input-triggered ASA. Let us recall its definition.

Definition 3 (Definition 4 in [2]): Let $\Pi = (K, \mathcal{E}, \mathcal{D})$ be a subversion relative to the symmetric encryption scheme $\Pi = (K, \mathcal{E}, \mathcal{D})$. Assume that the game $\text{SURV}_2$ and the game $\text{DETECT}$ are defined as in Fig. 3. A scheme $\Pi$ is said to be $\epsilon$-subversion resistant iff

$$\exists \mathcal{U}(B, \Pi) : \text{Adv}^{\text{SURV}}_{\Pi, \mathcal{U}}(B) \leq \epsilon$$

for all $\epsilon \in [0, 1]$.

DFP showed that $\epsilon$-subversion resistance is the best definition for preventing ASAs, when $\delta_2$ in Definition 2 is negligible but is not 0. In Theorem 2 of [2], they further claimed that the BPR unique ciphertext scheme in [1] is subversion resistant in terms of Definition 3. The definition of the BPR unique ciphertext scheme can be rewritten as follows.

Definition 4 (Definition 5 in [2]): Let $\Pi, \mathcal{K}, \mathcal{M}$ and $\mathcal{A}$ be same as those in Definition 1. A symmetric encryption scheme $\Pi = (K, \mathcal{E}, \mathcal{D})$ is said to have unique ciphertexts if:

1. $\Pi$ has perfect correctness and
2. for all $l \in \mathbb{N}$, all $K$, all $M$ and all $A$, there exists exactly one ciphertext vector $C$ such that:

$$(M, \sigma_l) \leftarrow D_K(C, A, \epsilon) \text{ for some } \sigma_l.$$  

In Theorem 2 of [2], a detection test $\mathcal{U}$ as Fig. 7 was proposed to prove the subversion resistance of the BPR unique ciphertext scheme in Definition 4. That is to say, for the BPR unique ciphertext scheme $\Pi$, any subversion $\Pi$, and any adversary $B$, the detection test $\mathcal{U}$ described in Fig. 7 achieves

$$\text{Adv}^{\text{SURV}}_{\Pi, \mathcal{U}}(B) \leq \text{Adv}^{\text{DETECT}}_{\Pi, \mathcal{U}}(B, \mathcal{U}).$$

The DFP surveillance game $\text{SURV}_2$ to the surveillance game $\text{SURV}$ as Fig. 8. In our surveillance game $\text{SURV}_2$, $B$ wins the surveillance game, when $B$ correctly outputs either the random bit $b$ or the encryption key $K$. We can further define the surveillance advantage of $B$ as follows.

$$\text{Adv}^{\text{SURV}}_{\Pi, \mathcal{U}}(B) = 2\Pr[\text{SURV}_2^{\mathcal{U}} = \text{true}] - 1.$$  

Here, we can see that $\text{Adv}^{\text{SURV}}_{\Pi, \mathcal{U}}(B)$ is 1, when BB can correctly output the encryption key $K$ of the encryption algorithm $E$. Otherwise, our proposed definition is same as the corresponding DFP definition. In fact, it must be pointed out that the surveillance game $\text{SURV}$ in the BPR model [1] also suffers from the similar flaw. In addition, the detection advantage of $\mathcal{U}$ is redefined by

$$\text{Adv}^{\text{DETECT}}_{\Pi, \mathcal{U}}(B, \mathcal{U}) = 2\Pr[\text{DETECT}_2^{B, \mathcal{U}} = \text{true}] - 1.$$  

Theorem 3: Let $\Pi = (K, \mathcal{E}, \mathcal{D})$ be the unique ciphertext symmetric encryption scheme as in Definition 4. Let $\Pi = (K, \mathcal{E}, \mathcal{D})$ defined in Fig. 4 be a subversion relative to $\Pi$. Assume that the game $\text{SURV}_2$ and the game $\text{DETECT}_2$ are defined as in Fig. 8. Assume that the detection test $\mathcal{U}$ is as Fig. 7. We have

$$\text{Adv}^{\text{SURV}}_{\Pi, \mathcal{U}}(B) > \text{Adv}^{\text{DETECT}}_{\Pi, \mathcal{U}}(B, \mathcal{U}).$$

Proof 3: Consider the game $\text{SURV}_2$ in Fig. 8. Given the subversion key $\tilde{K}$, $B$ has access to the key generation oracle $\text{Key}$ and the encryption oracle $\text{Enc}$. Here, the oracle $\text{Enc}$ invokes $\mathcal{E}$ described in Fig. 4 or $\mathcal{E}$ described in Definition 4, when the random bit $b$ is 0 or 1. $B$ makes a sequence of queries $(M = [M_1, \ldots, M_l], A = [A_1, \ldots, A_l])$ to the oracle $\text{Enc}$. The oracle $\text{Enc}$ then returns the corresponding $C = [C_1, \ldots, C_l]$. $B$ may call $\mathcal{E}$ as in Fig. 4 to recover the encryption key $K$. Let $E_3$ denote the event that $M' \neq M$, where $(M', \sigma_i) \leftarrow D_K(C, A, \epsilon)$. Let $E_4$ denote the event that $B$ believes in the existence of $M'$ and $M' = M$. According to our proposed ASA in Fig. 4, it requires the subverted ciphertext be exactly equal to the real unique ciphertext, i.e., $\mathcal{E}_K(M_i, A_i, \sigma_i) = \mathcal{E}_K(K, A_i, \sigma_i)$ for some $1 \leq i \leq l$, when $\text{pubinform} = \text{PUBINFORM}$. Clearly, the probability of the event $E_4$ is not 0. Let $E_5$ denote the event that $M' = M$ and the subverted ciphertext is not equal to the real unique ciphertext when $\text{pubinform} = \text{PUBINFORM}$. According to Eq. (6), the surveillance advantage can be expressed as

$$\text{Adv}^{\text{SURV}}_{\Pi, \mathcal{U}}(B) = 2\Pr[\text{SURV}_2^{\mathcal{U}} = \text{true}] - 1$$

If the event $E_3$ happens, $B$ knows that the oracle $\text{Enc}$ must use $\mathcal{E}$ because $\Pi$ in Definition 4 is perfectly correct. Hence, $\Pr[\text{SURV}_2^{\mathcal{U}} = \text{true}] = 1$. If the event $E_4$ occurs, $B$ should win the game $\text{SURV}_2$, because $\mathcal{U}$ can recover

VOLUME 8, 2020

175633
we further have the encryption key $K$ by its subversion key $\hat{K}$. Hence, $\Pr[\text{SURV}_2^{\hat{K}}] = \text{true}[E_3] = 1$. If the event $E_3$ happens, $B$ has no information about the random bit $b$ during the game $\text{SURV}_2$. Thus, $\Pr[\text{SURV}_2^{\hat{K}}] = \text{true}[E_3] = 1/2$. Therefore, we further have

$$\text{Adv}_{\text{SURV}_2^{\hat{K},\tilde{f}_1}}(B) = 2\Pr[E_3] + 2\Pr[E_4] + \frac{1}{2} \Pr[E_5] - 1 = \Pr[E_3] + \Pr[E_4].$$

Consider the game $\text{DETECT}_2$ in Fig. 8. The transcript $T$ records $K$ and $(M, A, C)$ of $B$ during the game $\text{SURV}_2$. Let $E_3$ denote the complementary event of $E_3$, i.e., $M' = M$, where $(M', \sigma) \leftarrow \hat{D}_K(C, A, \varepsilon)$. According to Eq. (7), we have

$$\text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B, U) = 2\Pr[\text{DETECT}_2^{\hat{K},\tilde{f}_1}] = \text{true}[E_3] \Pr[E_3] + 2\Pr[\text{DETECT}_2^{\hat{K},\tilde{f}_1}] = \Pr[E_3] - 1.\text{\text{Adv}}$$

For the detection test $U$ as in Fig. 7, we can calculate the detection advantage of $U$ with respect to $B$. We know that $U$ guesses either $b = 0$ ($M' \neq M$) or $b = 1$ ($M' = M$). As before, if $E_3$ occurs, $E_6$ exists and $b = 0$ due to the perfect correctness of $\hat{\Pi}$. Hence, $\Pr[\text{DETECT}_2^{\hat{K},\tilde{f}_1}] = \text{true}[E_3] = 1$. On the other side, when $E_3$ occurs, $U$ has no information about the random bit $b$ because $\hat{\Pi}$ has the unique ciphertext feature and $M' = M$. Thus, $\Pr[\text{DETECT}_2^{\hat{K},\tilde{f}_1}] = \text{true}[E_3] = 1/2$. We can therefore compute

$$\text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B, U) = 2\Pr[\text{DETECT}_2^{\hat{K},\tilde{f}_1}] = \text{true}[E_3] \Pr[E_3] + 2\Pr[\text{DETECT}_2^{\hat{K},\tilde{f}_1}] = \Pr[E_3] - 1.$$

We now have

$$\text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B) = \Pr[E_3] + 2\Pr[E_4] > \Pr[E_3] = \text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B, U).$$

This theorem shows that according to Definition 3, the detection test $U$ as in Fig. 7 cannot exclude the subversion $\hat{\Pi}$ as in Fig. 4 of the BPR unique ciphertext scheme $\Pi$, i.e., $\text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B) > \text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B, U)$, thereby reaching a contradiction. In addition, we can modify the input-triggered ASA to get a counterexample in Fig. 8 of [2]. That is, $E_6$ returns $(\hat{\varepsilon}(\hat{K}, K, A, \sigma), \sigma)$ instead of $(\hat{C})|\hat{A}, \sigma)$, when the predicate $R$ is true. And, BB with the subversion key $\hat{K}$ can recover the encryption key $K$ from $\hat{\varepsilon}(\hat{K}, K, A, \sigma)$. According to Definition 3, we claim that the modified input-triggered ASA also violates the subversion resistance of the BPR unique ciphertext scheme $\Pi$. The proof is the same as that of Theorem 3 and is therefore omitted.

The BPR unique ciphertext scheme still does not satisfy the DFP subversion resistance definition, when its subversion is allowed to have the trigger. We therefore need to improve the DFP subversion resistance definition to capture all existing trigger ASAs. Our improved definition is presented as follows.

**Definition 5 (Subversion Resistance):** Let $\tilde{\Pi} = (\hat{K}, \hat{\varepsilon}, \hat{\hat{L}})$ be any subversion relative to the symmetric encryption scheme $\Pi = (K, \varepsilon, D)$. Assume that the game $\text{SURV}_2$ and the game $\text{DETECT}_2$ are defined as in Fig. 8. We say $\Pi$ is subversion resistant if:

$$\exists U \forall (B, \tilde{\Pi}) : \text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B) - \text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B, U)$$

is a negligible value.

**Remark 4 (Analysis of Definition 5):** When $\Pi$ is a subversion resistant scheme, our improved definition does not strictly require $\text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B) \leq \text{Adv}_{\text{DETECT}_2^{\hat{K},\tilde{f}_1}}(B, U)$ for
any subversion $\tilde{\Pi}$ and any adversary $B$. Instead, it only requires $\text{Adv}_{\Pi, \tilde{\Pi}}^{\text{SURV}2}(B) - \text{Adv}_{\Pi, \tilde{\Pi}}^{\text{DET}}(B, U)$ to be negligible for any subversion $\tilde{\Pi}$ and any adversary $B$. We actually relax the DFP subversion resistance definition. Nevertheless, if $\text{Adv}_{\Pi, \tilde{\Pi}}^{\text{SURV}2}(B)$ is non-negligible, our improved definition assures $\text{Adv}_{\Pi, \tilde{\Pi}}^{\text{DET}}(B, U)$ is also non-negligible. Hence, our improved definition means that the detection test $U$ in the game $\text{DET}^{\text{ECT}2}$ of Fig. 8 should always be able to find the subversion behavior of $B$, once $B$ has efficiently subverted $\tilde{\Pi}$.

In the following, we show that the BPR unique ciphertext scheme is subversion resistant under Definition 5.

**Theorem 4**: Let $\Pi = (\mathcal{K}, \mathcal{E}, \mathcal{D})$ be the unique ciphertext symmetric encryption scheme as in Definition 4. Let $\tilde{\Pi} = (\tilde{\mathcal{K}}, \tilde{\mathcal{E}}, \tilde{\mathcal{D}})$ be any subversion relative to $\Pi$. Assume that the game $\text{SURV}2$ and the game $\text{DET}^{\text{ECT}2}$ are defined as in Fig. 8. Assume that the detection test $U$ is as Fig. 7. We have

$$\forall (B, \tilde{\Pi}) : \text{Adv}_{\Pi, \tilde{\Pi}}^{\text{SURV}2}(B) - \text{Adv}_{\Pi, \tilde{\Pi}}^{\text{DET}}(B, U) \text{ is a negligible value.}$$

**Proof 4**: Fix a subversion $\tilde{\Pi} = (\tilde{\mathcal{K}}, \tilde{\mathcal{E}}, \tilde{\mathcal{D}})$ and the corresponding adversary $B$ in the game $\text{SURV}2$ of Fig. 8. Let $E_3, \tilde{E}_3, E_4$, and $E_5$ respectively denote the same events as in Theorem 3. Actually, $E_3$, $E_4$, and $E_5$ cover all possible surveillance cases, after $B$ queries to the encryption oracle $\text{Enc}$ during the game $\text{SURV}2$. Similar to Theorem 3, we have

$$\text{Adv}_{\Pi, \tilde{\Pi}}^{\text{SURV}2}(B) = \text{Pr}[E_3] + \text{Pr}[E_4].$$

Consider the detection test $U$ as in Fig. 7. Based on the similar analysis in Theorem 3, the detection advantage of $U$ with respect to $B$ can be computed as

$$\text{Adv}_{\Pi, \tilde{\Pi}}^{\text{DET}}(B, U) = \text{Pr}[E_3].$$

Therefore, we have

$$\text{Adv}_{\Pi, \tilde{\Pi}}^{\text{SURV}2}(B) - \text{Adv}_{\Pi, \tilde{\Pi}}^{\text{DET}}(B, U) = \text{Pr}[E_3] + \text{Pr}[E_4] - \text{Pr}[E_3] = \text{Pr}[E_4].$$

Let us estimate $\text{Pr}[E_4]$. The event $E_4$ requires that $B$ recover the encryption key $K$ and $M' = M$. We argue that $\text{Pr}[E_4]$ is a negligible value by contradiction. Assume that $\text{Pr}[E_4]$ in the game $\text{SURV}2$ is non-negligible. We note that the BPR unique ciphertext scheme $\Pi$ should be a $v$-private scheme. That is, for the game PRIV in Fig. 5 and any adversary $A$, the advantage $\text{Adv}_{\Pi}^{\text{PRIV}}(A)$ defined in Eq. (2) is a negligible value $v$. However, the event $E_4$ directly constructs an adversary $A$, i.e., $B$, to break the BPR unique ciphertext scheme $\Pi$ in the game PRIV. Its advantage $\text{Adv}_{\Pi}^{\text{PRIV}}(B)$ is $\text{Pr}[E_4]$, which is not a negligible value. This contradicts the assumption of $\text{Pr}[E_4]$.

The above arguments conclude that $\text{Adv}_{\Pi, \tilde{\Pi}}^{\text{SURV}2}(B) - \text{Adv}_{\Pi, \tilde{\Pi}}^{\text{DET}}(B, U)$ is negligible. ■

**VI. CONCLUSION**

Our work dedicates to better understanding of how far we can extend the boundary of ASAs on symmetric encryption schemes. Our proposed ASA is valid under the game SDET of the BJK security model. Comparatively, the game DETECT in the DFP security model is able to detect our proposed ASA. In the view of cryptographic engineering, we argue that the game DETECT is more expensive than the game SDET. SD in the game SDET can only experiment with the encryption oracle $\text{Enc}$. However, the game DETECT needs to record all oracle $\text{Enc}$ queries of $B$. We have also shown that the DFP subversion resistance definition is unreasonable and therefore improved their definition. It needs to be pointed out that the improved definition is merely suitable for detecting the trigger ASAs. We still do not know whether the improved definition can capture ASAs in [1], [3]. This problem needs to be investigated in the future. In practice, the probability of decryptability defined in Definition 2 is perhaps a small but non-negligible value. Hence, another future work is to explore new ASAs, define the security model, and design the symmetric encryption schemes under this new security assumption. In addition, we will also substitute the applied symmetric encryption algorithms under computer and network systems and evaluate the performance of our proposed ASA.

**ACKNOWLEDGMENT**

Part of this work was done in University of Wollongong (UOW), Australia. The authors would like to thank the editor and the reviewers for their valuable suggestions and comments.

**REFERENCES**

[1] M. Bellare, K. G. Paterson, and P. Rogaway, “Security of symmetric encryption against mass surveillance,” in *Proc. CRYPTO* (Lecture Notes in Computer Science), vol. 8616, J. A. Garay and R. Gennaro, Eds. Berlin, Germany: Springer, 2014, pp. 1–19.

[2] J. P. Degabriele, P. Farshim, and B. Poettering, “A more cautious approach to security against mass surveillance,” in *Proc. FSE* (Lecture Notes in Computer Science), vol. 9054, G. Leander, Ed. Berlin, Germany: Springer, 2015, pp. 579–598.

[3] M. Bellare, J. Jaeger, and D. Kane, “Mass-surveillance without the state: Strongly undetectable algorithm-substitution attacks,” in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2015, pp. 1431–1440.

[4] G. J. Simmons, “The prisoners’ problem and the subliminal channel,” in *Proc. CRYPTO* D. Chaum, Ed. Berlin, Germany: Springer, 1984, pp. 51–67.

[5] G. J. Simmons, “The subliminal channel and digital signature,” in *Proc. EUROCRYPT* (Lecture Notes in Computer Science), vol. 209, T. Beth, N. C. Cot, and I. Ingemarsson, Eds. Berlin, Germany: Springer, 1984, pp. 364–378.

[6] A. Young and M. Yung, “The dark side of cryptography or: Should we trust capstone?” in *Proc. CRYPTO* (Lecture Notes in Computer Science), vol. 1109, N. Koblitz, Ed. Berlin, Germany: Springer, 1996, pp. 89–103.

[7] A. Young and M. Yung, “Kleptography: Using cryptography against cryptography,” in *Proc. EUROCRYPT* (Lecture Notes in Computer Science), vol. 1233, W. Fumy, Ed. Berlin, Germany: Springer, 1997, pp. 62–74.

[8] A. Young and M. Yung, “Monkeys: Black-box symmetric ciphers designed for monopolizing keys,” in *Proc. FSE* (Lecture Notes in Computer Science), vol. 1372, S. Vaudenay, Ed. Berlin, Germany: Springer, 1998, pp. 112–133.

[9] V. Rijmen and B. Preneel, “A family of trapdoor ciphers,” in *Proc. FSE* (Lecture Notes in Computer Science), vol. 1267, E. Biham, Ed. Berlin, Germany: Springer, 1999, pp. 201–214.
[11] E.-J. Goh, D. Boneh, B. Pinkas, and P. Golle, “The design and implementation of protocol-based hidden key recovery,” in *Proc. ISC* (Lecture Notes in Computer Science), vol. 2851, L. Knudsen, Ed. Berlin, Germany: Springer, 2003, pp. 165–179.

[12] A. Waksman and S. Sethumadhavan, “Silencing hardware backdoors,” in *Proc. IEEE Symp. Security and Privacy*, May 2011, pp. 49–63.

[13] J. Ball, J. Borger, and G. Greenwald, “Revealed: How US and UK spy agencies defeat Internet privacy and security,” *Guardian*, Sep. 2013. [Online]. Available: http://www.theguardian.com/world/2013/sep/05/ NSA-gchq-encryption-codes-security

[14] G. Greenwald, *No Place to Hide: Edward Snowden, the NSA, and the US Surveillance State*. New York, NY, USA: Metropolitan Books/Henry Holt, 2015, p. 259.

[15] J. Baek and I. You, “Efficient substitution of symmetric encryption with random initialization vector,” *IEICE Trans. Inf. Syst.*, vol. 99, no. 4, pp. 1251–1254, 2016.

[16] M. Armour and B. Poettering, “Subverting decryption in AEAD,” in *Proc. IMACC* (Lecture Notes in Computer Science), vol. 11929, M. Albrecht, Ed. Cham, Switzerland: Springer, 2019, pp. 22–41.

[17] A. Russell, Q. Tang, M. Yang, and H.-S. Zhou, “Cliquotography: Clipping the power of kleptographic attacks,” IACR Cryptol. ePrint Arch., Tech. Rep. 695, 2015.

[18] Y. Dodis, C. Ganesh, A. Golovnev, A. Juels, and T. Ristenpart, “A formal treatment of backdoored pseudorandom generators,” in *Proc. EUROCRYPT* (Lecture Notes in Computer Science), vol. 9056, E. Oswald and M. Fischlin, Eds. Cham, Switzerland: Springer, 2015, pp. 101–126.

[19] I. Mironov and N. Stephens-Davidowitz, “Cryptographic reverse firewalls,” in *Proc. EUROCRYPT* (Lecture Notes in Computer Science), vol. 9057, E. Oswald and M. Fischlin, Eds. Cham, Switzerland: Springer, 2015, pp. 657–686.

[20] G. Ateniese, B. Magri, and D. Venturi, “Subversion-resilient signature schemes,” in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2015, pp. 364–375.

[21] Y. Dodis, I. Mironov, and N. Stephens-Davidowitz, “Message transmission with reverse firewalls—Secure communication on corrupted machines,” in *Proc. CRYPTO* (Lecture Notes in Computer Science), vol. 9814, M. Robshaw and J. Katz, Eds. Berlin, Germany: Springer, 2016, pp. 341–372.

[22] G. Ateniese, B. Magri, and D. Venturi, “Subversion-resilient signatures: Definitions, constructions and applications,” *Theor. Comput. Sci.*, vol. 820, pp. 91–122, Jun. 2020.

[23] C. Liu, R. Chen, Y. Wang, and Y. Wang, “Asymmetric subversion attacks on signature schemes,” in *Proc. ACISP* (Lecture Notes in Computer Science), vol. 10946, W. Susilo and G. Yang, Eds. Cham, Switzerland: Springer, 2018, pp. 376–395.

[24] Y. Wang, R. Chen, C. Liu, B. Wang, and Y. Wang, “Asymmetric subversion attacks on signature and identification schemes,” Pers. Ubiquit. Comput., to be published. Accessed: Aug. 23, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s00779-018-01193-x; doi: 10.1007/s00779-018-01193-x.

[25] S. Dziembowski, S. Faust, and F.-X. Standaert, “Private circuits III: Hardware trojan-resilience via testing amplification,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 142–153.

[26] G. Ateniese, A. Kayiaysi, B. Magrithm, Y. Tselekounisy, and D. Venturix, “Secure outsourcing of circuit manufacturing,” IACR Cryptol. ePrint Arch., Tech. Rep. 527, 2016.

[27] M. Bellare, D. Kane, and P. Rogaway, “Big-key symmetric encryption: Resisting key exfiltration,” in *Proc. CRYPTO* (Lecture Notes in Computer Science), vol. 9814, M. Robshaw and J. Katz, Eds. Berlin, Germany: Springer, 2016, pp. 373–402.

[28] M. Fischlin and S. Mazaheri, “Self-guarding cryptographic protocols against algorithm substitution attacks,” IACR Cryptol. ePrint Arch., Tech. Rep. 31, 2017.

[29] S. Bernt and M. Liškiewicz, “Algorithm substitution attacks from a steganographic perspective,” in *Proc. ACISP* (Lecture Notes in Computer Science), vol. 10769, M. Abdalla and R. Dahab, Eds. Cham, Switzerland: Springer, 2018, pp. 190–218.

[30] B. Auerbach, M. Bellare, and E. Kiltz, “Public-key encryption resistant to parameter subversion and its realization from efficiently-embeddable groups,” in *Proc. PKC* (Lecture Notes in Computer Science), vol. 10769, M. Abdalla and R. Dahab, Eds. Cham, Switzerland: Springer, 2018, pp. 348–377.