SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2016, Vol. 59 112101:1–112101:11
doi: 10.1007/s11432-015-5474-8

Discussion on the theoretical results of white-box cryptography

Tingting LIN, Xuejia LAI*, Weijia XUE & Geshi HUANG

Cryptography and Information Security Lab, Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200240*, China*

Received July 13, 2015; accepted September 26, 2015; published online May 24, 2016

Abstract White-box cryptography (WBC) aims to resist attacks from attackers who can control all the implementation details of cryptographic schemes. In 2009, Saxena et al. proposed a fundamental of white-box cryptography via the notion "white-box property" (WBP). Under this model, they proved that there do not exist obfuscators that can satisfy every security notion for a program (the negative result). On the other hand, they proved that there exists an obfuscator satisfying WBP for some security notion (the positive result). These contributions provide us a general cognition of WBC, which is big progress for the theoretical research. To better understand them, we make discussion on each result and achieve some new results. For the negative result, we prove that insufficiently secure obfuscator is the real cause of the negative result. We point out that the security of a white-box scheme cannot be guaranteed if it is instantiated by a less secure obfuscator, since the obfuscator used in their proof does not satisfy the "Virtual Black-box Property" with auxiliary input. From our proof, we also conclude that the notion WBP is equal to "Virtual Black-box Property with auxiliary input". For the positive result, we prove that security notion under black-box model should not be used in white-box context without any modification; although the positive result is meaningful, it is unlikely to prove that an obfuscator satisfies WBP for IND-CPA, since the security notion "IND-CPA" is under black-box model, which has different adversary with WBP.

Keywords white-box, obfuscation, cryptography, IND-CPA, white-box property (WBP)

Citation Lin T T, Lai X J, Xue W J, et al. Discussion on the theoretical results of white-box cryptography. Sci China Inf Sci, 2016, 59(11): 112101, doi: 10.1007/s11432-015-5474-8

1 Introduction

Traditional cryptography is based on black-box attack model, which assumes the communication endpoint and operating environment of a cryptsystem are trusted, an adversary has only oracle access to the cryptosystem. In the past several decades, traditional cryptography has made rapid development in promoting technological innovation [1–3], but meanwhile, cryptanalysis technologies have also undergone tremendous changes, like white-box attack. White-box attack is a more powerful attack type and different with traditional attacks. In [4], Chow et al. first presented the notion of white-box attack context (WBAC), which assumed the adversary has high privilege in a host and has full control of the

* Corresponding author (email: lai-xj@cs.sjtu.edu.cn)

cryptographic software; it can observe all the dynamic execution and internal running details, whenever it wants and with whoever it can find.

Unfortunately, white-box attack becomes more popular with mass applications of cryptography, such as mobile agent, digital content distribution, cloud computing. Inevitably, such cryptographic applications are often implemented in an un-trusted endpoint, the execution details such as memory states, CPU calls, dynamic data can be inspected or intercepted by the adversary, traditional security model which assumes that the adversary has only access to the functionality of a cryptosystem cannot counter such attacks. White-box cryptography (WBC) was proposed to address this problem—it aims to remain "secure" even the implementation of the cryptographic software is under white-box attacks.

In the year of 2002, Chow et al. proposed practical white-box AES implementation scheme in [4] and white-box DES implementation scheme in [5] by using look-up tables respectively. In 2009 and 2011, two variants of Chow et al.'s scheme were presented by Xiao and Lai in [6], and by Karroumi in [7] respectively. Another attempt to design white-box AES was presented by Bringer et al. in [8] by using the method of perturbations. In 2009 and 2015, two white-box implementations of SMS4 were proposed by Xiao et al. in [9], and by Shi et al. in [10] respectively. Unfortunately, almost all the aforementioned schemes were lack of formal security proof and proved to be un-secure by some cryptanalysis [11–19]. In 2014, Alex Biryukov et al. presented several general white-box design methods with ASASA structure and specified their security, but one of the cryptosystems was attacked by Gilbert et al. in [20].

On the other hand, the basic theoretical research of WBC developed slowly, only a few studies were published. In 2008, Herzberg et al. presented the notion of white-box remote program execution (WBRPE) in [21] and proved that it was possible to execute a program in remote environment with some security specifications. WBRPE can be viewed as an extension of WBC.

In 2009, Saxena et al. presented a fundamental of white-box cryptography via the notion "white-box property" (WBP) in [22], they also proposed two conclusions about WBC. One of them is a negative result which shows that for the non-approximately-learnable family, no obfuscators can satisfy white-box property for all security notions; the other is a positive result which shows that if bilinear Diffie-Hellman assumption holds and a hash function is a random oracle, there exists an obfuscator satisfying WBP for the security notion IND-CPA.

Compared with Herzberg et al.'s work, Saxena et al. studied WBC specifically. Their work aimed at the underlying problems in WBC and proved their points on some critical results for the first time. Their work also outlined a theoretical foundation of WBC and have important implications for white-box cryptography. But there still exists something unclear in proofs of the two results. For better understanding and application, we rethink the results and make discussion on each of them. We also achieve some new deep-going results.

It should be noticed at first that obfuscator is used as a method to instantiate white-box scheme in [22], which has been specified in the first section of the full version [23] of [22]. That is to say, for a scheme M, the obfuscation $\mathcal{O}(M)$ is the white-box version of M, so we use obfuscation and white-box scheme indiscriminately in the rest parts of the paper.

For the negative result, we point out that the impossibility result is in fact caused by using an obfuscator which does not satisfy the "Virtual Black-box Property" with auxiliary input, which is the strongest security requirement for obfuscator as far as we know. In other words, insufficiently secure obfuscator cannot be used to instantiate all ideal white-box scheme; but how secure an obfuscator is good for white-box scheme is an object of the future research.

For the positive result, we show an overlooked fact. We indicate that the security notion "IND-CPA" is not a notion under the white-box attack assumption, but a notion under "black-box" attack assumption (i.e. the adversary has only oracle-access to the encryption scheme). Hence, it is inappropriate to say that the obfuscator satisfies the WBP for "IND-CPA". By analogy, security notions under black-box model such as CPA and CCA should not be used in white-box context directly. White-box cryptography ought to have its own security requirement, which is also another object of the future research.

Our work is crucial to the central theoretical research of WBC. Under the security notion WBP, our first result is a complement, which shows not only what obfuscator should not be used to instantiate white-box scheme, but also the relationship between WBP and "Virtual Black-box Property with auxiliary input", that is, they are equivalent if obfuscator is used as a method to instantiate white-box scheme. Our second result is in fact a warning of the usage of WBP and "black-box security notion", we clarify the difference between white-box adversary and black-box adversary, which is prone to be neglected.

2 Preliminaries

2.1 Symbols and notations

In order for keeping consistent, we use the similar notations as [22]

— $\mathbb{P}$ is the collection of all polynomials which have coefficients in $[0, \infty]$.

— $\mathbb{PPT}$ is the collection of all probabilistic polynomial time (PPT) algorithms.

— $\mathbb{TMF}$ is the collection of all turing machine family (TMFs), which are Turing Machines having a key read tape and an input read tape.

— $\mathcal{K}_M^l$ is key space of a TMF M, the length of the key is l.

— $\mathcal{I}_M^l$ is the input space of $M[k]$ for a fixed key $k \in \mathcal{K}_M^l$.

— M is a polynomial turing machine family (PTMF) if M is a TMF at first and $M[k](x)$ terminates in at most $p(l)$ steps for a polynomial $p \in \mathbb{P}$, any $k \in \mathcal{K}_M^l$ and any $x \in \mathcal{I}_M^l$. $\mathbb{PTMF}$ is the collection of all the PTMFs.

— A mapping h ($h : \mathbb{N} \to \mathbb{R}$) is said to be negligible in x if for any $p \in \mathbb{P}$, there exists a $x' \in \mathbb{N}$, such that $\forall x > x' : h(x) < 1/p(x)$.

2.2 Approx.Learnable Family

The notion of learnable function has been introduced in [24–27]. In [22], the authors gave a definition of Approx.Learnable Family, we repeat it as follows:

Definition 1 (Approx.Learnable Function). A function F is approximately learnable if there exists a circuit C that can output a polynomial sized circuit C_F by making limited times oracle queries, such that C_F is functionally equivalent with F on almost all of the inputs:

$$\Pr[x \overset{R}{\leftarrow} \mathcal{I}_F; C_F \leftarrow C^F(1^n) : C_F(x) = F(x)] \geqslant 1/\text{negl}(l),$$

where I_F is the input-space of F, negl is a negligible function and l is the security parameter. The collection of all the approximately learnable functions is denoted by $\mathbb{ALF}$.

2.3 Obfuscation

Generally speaking, an obfuscator is a "compiler" that turns a program (a circuit) Q into a new program $\mathcal{O}(Q)$, its purpose is to obscure the program and make it "unintelligible" in some sense. The first theoretical contributions towards obfuscation were presented by Hada [28] and Barak et al. [29]. Since then many negative and positive results of obfuscation have been achieved.

The negative results of obfuscation were mainly caused by strong definitions. In [29], a strong definition of "Virtual Black-box" obfuscation was presented, which essentially required that one cannot learn more information from the obfuscated program than one can learn from black-box (oracle) access to that program. In [29], Barak et al. also ruled out the existence of a general obfuscator by showing that there exist a family of functions that cannot be obfuscated in the sense of "Virtual Black-box". However, their work still left open the possibility that some specific functions and a lot of cryptsystems might be "Virtual Black-box" obfuscated. Afterwards, many results of "Virtual Black-box" obfuscation were presented in [26, 27, 30–32].

In 2005, the definition of "Virtual Black-box" obfuscation was strengthened by Goldwasser et al. [33], thus a super-strong definition of "Virtual Black-box" obfuscation with regard to auxiliary input was presented. This definition required that for any auxiliary input ai, one cannot learn more information from the obfuscated program and ai than one can learn from black-box (oracle) access to that program and ai:

Definition 2 ("Virtual Black-box" obfuscation with auxiliary input)**.** $\mathcal{O} \in \mathbb{PPT}$ is said to be an obfuscator for $M \in \mathbb{PTMF}$ if it satisfies the following conditions:

1. (Functionality) For any input length l, $\exists i \in \{0,1\}^l$, $\Pr[\mathcal{O}(M)(i) \neq M(i)] \leqslant \text{negl}(l)$, where $\text{negl}(\cdot)$ is a negligible function.

2. (Polynomial-Slowdown) For a polynomial $h(l)$, s.t. $|\mathcal{O}(M)| \leqslant h(|M|)$.

3. ("Virtual Black-box Property" with auxiliary input) For any probabilistic polynomial-time adversary A, any predicate π, any polynomial $p(\cdot)$, any auxiliary input 'ai' of size $p(l)$, there is a probabilistic polynomial-time simulator S, such that $|\Pr[A(\mathcal{O}(M), ai) = \pi(M, ai)] - \Pr[S^M(1^l, ai) = \pi(M, ai)]| \leqslant \text{negl}(l)$, where auxiliary input ai may or may not depend on M.

Considering auxiliary input is important when we use obfuscation in a larger protocol or system. In general, a program is often implemented in a larger system, it has to coexist with other related components. Hence, an adversary can always obtain some information from such components when it runs the program, such as the functionality of a system library function, priori information of a protocol. Information like this will be called auxiliary inputs of the adversary. With this definition, the positive results [26, 27] based on the "Virtual Black-box" obfuscation were proved to be not really true [33].

Recently, many positive results were proposed by using relaxed definition of obfuscation: indistinguishability obfuscation (iO). Namely, iO requires only that it is computationally hard to distinguish the obfuscation of a circuit C_0 from that of another circuit C_1, where C_0 and C_1 are functionally equivalent and of the same size. In [34], Garg et al. introduced a candidate iO construction for all circuits. Subsequently, a number of applications of iO [35–37] have appeared to show that iO may be the "best possible obfuscation" as it was claimed in [38]. In 2012, Barak et al. [39] extended the notion iO into a stronger (but still weak) notion differing-inputs obfuscation (diO), which requires that the difficulty of finding an input x on which $C_0(x) \neq C_1(x)$ implies it is computationally hard to distinguish $\text{diO}(C_0)$ and $\text{diO}(C_1)$. Afterwards, both of the two definitions were strengthened by adding the condition "with auxiliary input" (but still weaker than Definition 2), e.g. indistinguishability obfuscation with auxiliary input and differing-inputs obfuscation with auxiliary input, each one corresponded to several positive and negative results [40–42].

2.4 White-box attack

According to Chow et al. [4], the white-box attack context (WBAC) can be concluded as follows:

(1) Attack software is fully-privileged on a host and has complete visibility of the implementation of a cryptographic software.

(2) Dynamic execution results of the cryptographic software can be obtained.

(3) Implementation details of the cryptographic software can be altered at will.

In fact, white-box attack also includes reverse engineering, malicious host attack, side-channel attack and so on. With white-box attack, an adversary can view the information in cache or memory, trace program instructions, halt execution at any desired point and change program code or memory, etc. The adversary can do all the things at anytime with anyone's help. Thus we could refer to the white-box attack model as the worst-case model, the white-box adversary as the strongest adversary.

3 Revisit the definition of white-box property

The notion of WBP is the base of other results in [22], it is defined with a black-box game (BBgame) and a white-box game (WBgame). But first of all, we need to introduce "security notion (sn)":

Definition 3 (Security Notion (sn), Definition 5 of [22])**.** sn is defined as a 5-tuple $(n, l, \boldsymbol{M}, \text{Extr}, \text{Win})$, where n is a natural number, l is a security parameter, $\boldsymbol{M} = (M_1, M_2, \ldots, M_n)$ is an n-tuple of PTMFs, Extr: $\{0,1\}^l \to \times_{i=1}^{n} \mathcal{K}_{M_i}$ and Win: $\{0,1\}^* \to \{0,1\}$ are both TMs. $\mathbb{SN}$ is defined as the collection of all security notions.

A security notion is often used with games, we will explain it in the following two games. For more details, please refer to [22].

BBgame is in fact the traditional experiment used in security proof, it takes $(1^l, \mathrm{sn}, r)$ as input, then parses sn as $(n, l, \boldsymbol{M} = (M_1, M_2, \ldots, M_n), \mathrm{Extr}, \mathrm{Win})$, where Extr will extract parameters for $M_1, M_2, \ldots, M_n$, Win will provide its decision 0 or 1 as the output of this game; especially in the game, the adversary is only given oracle access to $\boldsymbol{M}$. It is denoted by

$$\mathsf{AdvBB}_A^{\mathrm{sn}}(l) = \Pr[r \overset{R}{\leftarrow} \{0,1\}^l : \mathrm{BBgame}_A(1^l, \mathrm{sn}, r) = 1],$$

the adversary's advantage in BBgame.

$\underline{\mathrm{BBgame}_A(1^l, \mathrm{sn}, r)}$:

— sn is parsed as five items $n, l, \boldsymbol{M}, \mathrm{Extr}, \mathrm{Win}$;

— $\boldsymbol{M}$ is parsed as n items $M_1, M_2, \ldots, M_n$;

— Extract keys $k_1, k_2, \ldots, k_n$;

— The adversary A is given security parameter k, sn and oracle-access to $M_1[k_1], M_2[k_2], \ldots, M_n[k_n]$;

— A then outputs d;

— The output of the game is $\mathrm{Win}(r, \mathrm{Queries}, d)$,

where Queries is a collection of all 4-tuples: $(t_s, s, \mathrm{in}_s, \mathrm{out}_s)$, the four items denote time, sequence number of oracle, input and output of a query respectively.

Before defining WBgame, it is necessary to formalize "obfuscatable". In fact, M_i is obfuscatable w.r.t sn means that M_i satisfies sn and there exists an obfuscation $\mathcal{O}(M_i)$ can be used to replace the oracle of M_i when it is queried in a game w.r.t sn, so Win's output remains unchanged whether M_i's oracle exists or not.

Definition 4 (Definition 7 in [22])**.** M_i is said to be obfuscatable w.r.t sn (denote by $M_i \in_{\mathrm{obf}}$ sn) if $\mathrm{Win}(r, \mathrm{Queries}, d) = \mathrm{Win}(r, \mathrm{Queries}(i), d)$ for any r, Queries, d, where $\mathrm{Queries}(i)$ is a set: $\{(t_s, s, \mathrm{in}_s, \mathrm{out}_s) \,|(t_s, s, \mathrm{in}_s, \mathrm{out}_s) \in \mathrm{Queries} \wedge s \neq i\}$.

In WBgame, the input and output are similar as those in BBgame, the only difference is that the adversary is not only given oracle access to $\boldsymbol{M}$, but also given white-box access to $\mathcal{O}(M_i)$. That is, the adversary totally controls the running details of $\mathcal{O}(M_i)$, including attacks like observation, interception, debugging, tampering, reverse-engineering, etc. It is denoted by $\mathsf{AdvWB}_{A,\mathcal{O},M_i}^{\mathrm{sn}}(l) = \Pr[r \overset{R}{\leftarrow} \{0,1\}^l : \mathrm{WBgame}_{A,\mathcal{O},M_i}(1^l, \mathrm{sn}, r) = 1]$ the adversary's advantage in WBgame.

$\underline{\mathrm{WBgame}_{A,\mathcal{O},M_i}(1^l, \mathrm{sn}, r)}$:

— sn is parsed as five items $n, l, \boldsymbol{M}, \mathrm{Extr}, \mathrm{Win}$;

— $\boldsymbol{M}$ is parsed as n items $M_1, M_2, \ldots, M_n$;

— Extract keys $k_1, k_2, \ldots, k_n$;

— The adversary A is given security parameter k, sn, M_i's obfuscation $\mathcal{O}(M_i, k_i)$ and oracle-access to $M_1[k_1], M_2[k_2], \ldots, M_n[k_n]$;

— A then outputs d;

— The output of the game is $\mathrm{Win}(r, \mathrm{Queries}, d)$,

where A is given $\mathcal{O}(M_i, k_i)$ means that A has white-box access to it.

Definition 5 (White-box Property (WBP), Definition 10 of [22])**.** $\mathcal{O}$ is said to satisfy white-box property w.r.t (M, sn) if

$$|\max(\mathsf{AdvWB}_{A,\mathcal{O},M}^{\mathrm{sn}}(l)) - \max(\mathsf{AdvBB}_A^{\mathrm{sn}}(l))| \leqslant \mathrm{negl}(|l|),$$

for any $(\mathcal{O}, M, \mathrm{sn}) \in \mathbb{PPT} \times \mathbb{PTMF} \times \mathbb{SN}$.

Definition 6 (universal white-box property (UWBP), Definition 11 of [22])**.** $\mathcal{O}$ is said to satisfy universal white-box property w.r.t M if for any $\mathrm{sn} \in \mathbb{SN}$, either M is not obfuscatable or $\mathcal{O}$ satisfies WBP w.r.t (M, sn), for any $(\mathcal{O}, M) \in \mathbb{PPT} \times \mathbb{PTMF}$.

4 Discussion on the negative result

4.1 The negative result

In [22], based on the definitions of WBP and UWBP, it is proved in negative result that there does not exist an obfuscator that satisfies UWBP for a "special" family—non-approximately-learnable family. The proof is completed by constructing an algorithm and showing the obfuscation $\mathcal{O}(M)$ (i.e. a white-box scheme) fails to satisfy an sn stated in the algorithm.

Theorem 1 (Theorem 1 of [22])**.** For any $(\mathcal{O}, M)$, where $\mathcal{O} \in \mathbb{PPT}$ and $M \in \mathbb{PTMF}\backslash\mathbb{ALF}$, there exist $\mathrm{sn} \in \mathbb{SN}$, s.t. $M \in_{\mathrm{obf}} \mathrm{sn}$ whereas $\mathcal{O}$ cannot satisfy WBP w.r.t (Q, sn).

In the proof of this theorem, sn is defined as guess-u $= (2, l, \boldsymbol{M}, \mathrm{Extr}, \mathrm{Win})$, where $\boldsymbol{M} = (M, M_1)$, the algorithm is as follows:

Algorithm 1:

```
— Function $M_1[k_1](Y_1)$:
        $k_1$ is parsed as $(k, u, x)$;
        if $(Y_1(x) = M[k](x))$
            then output $u$;
        else output 0;
— Function Extr$(r)$:
        $r$ is parsed as three items $(k, u, x)$;
        $k_1$ is set to be $k, u, x$;
        output $k, k_1$;
— Function Win$(r, \mathrm{Queries}, d)$:
        $r$ is parsed as three items $(k, u, x)$;
        if $(d = u)$ and $M_1[k_1]$ was queried less than or equal to once
            then output 1;
        else output 0.
```

Because $M \notin \mathbb{ALF}$, according to Definition 1, it means the probability of outputting the function C_M satisfying $C_M(x) = M[k](x)$ just by using learning machine C and oracle-access to $M[k]$ is negligible. That is to say in Algorithm 1 it is hard to find a Y_1 by oracle-access to $M[k]$, this is the case of BBgame w.r.t Algorithm 1. Therefore, the advantage of the adversary A in BBgame is negligible:

$$\forall A \in \mathbb{PPT} : 0 \leqslant \mathsf{AdvBB}_{A}^{\text{guess-}u}(l) < \mathrm{negl}(l).$$

However, in the WBgame, A is given white-box access to $\mathcal{O}(M)$ and has its full control of $\mathcal{O}(M)$, therefore with a non-negligible probability, it can input $\mathcal{O}(M)$ to $M_1[k_1]$ to get u since $\mathcal{O}(M)$ is functionally equivalent to $M[k]$:

$$\forall A \in \mathbb{PPT} : 1 \geqslant \mathsf{AdvWB}_{A,\mathcal{O},M}^{\text{guess-}u}(l) > 1 - \mathrm{negl}(l).$$

Hence, $|\max(\mathsf{AdvWB}_{A,\mathcal{O},M}^{\mathrm{sn}}(l)) - \max(\mathsf{AdvBB}_{A}^{\mathrm{sn}}(l))| > 1 - \mathrm{negl}(l)$, $\mathcal{O}$ cannot satisfy WBP for (M, sn).

4.2 The real cause of the negative result

First of all, we need to note that obfuscator $\mathcal{O}$ is in fact used as a method to instantiate WBC in Theorem 1, which means $\mathcal{O}(M)$ is the white-box scheme of M, and the "obfuscation adversary" is equivalent to white-box adversary.

In the following, we will use a theorem to show that an insufficient secure obfuscator will lead to an insecure white-box scheme in the sense of WBP. Also, by this theorem, we point out that the obfuscator $\mathcal{O}$ used in Theorem 1 does not satisfy the "Virtual Black-box Property" with auxiliary input and this is the inherent reason of why the negative result holds in Theorem 1.

Theorem 2. There exists such white-box scheme which fails to satisfy WBP for a security notion is caused by the fact that the obfuscator $\mathcal{O}$ used to instantiate it does not possess the "Virtual Black-box Property with auxiliary input".

Proof. We use Algorithm 1 as the example.

In Theorem 1, security notion guess-u is defined with the system $\boldsymbol{M}(M,M_1)$, M and M_1 co-exist in one system. According to Subsection 2.3, information from M_1 is the auxiliary inputs 'ai'.

On one hand, we say for a probabilistic polynomial time adversary A that is given $\mathcal{O}(M)$ and M_1, it is easy to get u. Since $\mathcal{O}(M)$ has the functionality of M, M_1 will output u when the adversary use $\mathcal{O}(M)$ as the input to M_1 (see the details of Algorithm 1):

$$\Pr[A(\mathcal{O}(M), M_1) = u] \geqslant 1 - \text{negl}(l).$$

But for any simulator S, it has only black-box access (oracle-access) to M, the probability of outputting u is equal to the probability of finding a Y_1 satisfying $Y_1(x) = M[k](x)$, which is equal to learning M. This probability is less than a negligible value since $M \notin \mathbb{ALF}$:

$$\Pr[S^M(1^l, M_1) = x] < \text{negl}(l),$$

then it can be deduced that

$$|\Pr[A(\mathcal{O}(M), M_1) = u] - \Pr[S^M(1^l, M_1) = u]| \geqslant 1 - \text{negl}(l).$$

So the obfuscator $\mathcal{O}$ used in Theorem 1 does not possess the "Virtual Black-box Property with auxiliary input".

On the other hand, assume the obfuscator $\mathcal{O}$ used in Theorem 1 possesses the "Virtual Black-box Property with auxiliary input", according to Definition 2, for any probabilistic polynomial-time adversary A, any input length l, there exists a probabilistic polynomial-time simulator S, such that

$$|\Pr[A(\mathcal{O}(M), M_1) = u] - \Pr[S^M(1^l, M_1) = u]| \leqslant \text{negl}(l).$$

Now we consider the security notion WBP. Firstly in BBgame, we "guess-u" by calling above simulator S and get

$$\mathsf{AdvBB}_A^{\text{guess-u}}(l) = \Pr[S^M(1^l, M_1) = u],$$

hence for all the security parameter l:

$$\begin{aligned}\max(\Pr[A(\mathcal{O}(M), M_1) = u]) - \text{negl}(l) &\leqslant \max(\mathsf{AdvBB}_A^{\text{guess-u}}(l)) \\ &\leqslant \max(\Pr[A(\mathcal{O}(M), M_1) = u]) + \text{negl}(l).\end{aligned} \tag{1}$$

Secondly, we are going to discuss the white-box adversary's advantage $\mathsf{AdvWB}_{A,\mathcal{O},M}^{\text{guess-u}}(l)$. Since the "obfuscation adversary" is equivalent to white-box adversary, for the same game "guess-u", the white-box adversary's best probability of winning is equal to "obfuscation adversary"'s best probability of winning, for all the security parameter l:

$$\max(\mathsf{AdvWB}_{A,\mathcal{O},M}^{\text{guess-u}}(l)) = \max(\Pr[A(\mathcal{O}(M), M_1) = u]). \tag{2}$$

With inequality (1) and equality (2), we can get

$$|\max(\mathsf{AdvWB}_{A,\mathcal{O},M}^{\text{sn}}(l)) - \max(\mathsf{AdvBB}_A^{\text{sn}}(l))| \leqslant \text{negl}(l),$$

which means if the obfuscator $\mathcal{O}$ used in Theorem 1 possesses the "Virtual Black-box Property with auxiliary input", the white-box scheme will satisfy WBP for the security notion "guess-u".

Remark 1. It should have been noticed that WBP is somewhat like "Virtual Black-box Property with auxiliary input", or WBC is somewhat like obfuscation. For the relation between WBC and obfuscation, there exist many viewpoints [11, 21, 23]. It is well accepted that obfuscation is a candidate method for WBC, WBC is a specific field of obfuscation. In fact, another interpretation of Theorem 2 is: if obfuscator is used as a method to instantiate white-box scheme and the "obfuscation adversary" is equivalent to white-box adversary, the notion WBP is equal to "Virtual Black-box Property with auxiliary input".

Remark 2. This conclusion suggests that it should be careful to use weak secure obfuscator to instantiate white-box scheme, since such obfuscator might be the direct reason that leads to insecure white-box scheme.

5 Discussion on the positive result

5.1 The positive result

In the positive result, a theorem is proposed to show that there exists an obfuscator $\mathcal{O}$ that can satisfy WBP for security notion IND-CPA. In the theorem, a symmetric encryption scheme $\mathcal{E} = (G, E, D)$ is defined as follows:

(1) The key generation G is defined as

$$sk = (\hat{e},\ F_1,\ F_2,\ q,\ H,\ a,\ x),$$

where $\hat{e} : F_1 \times F_1 \to F_2$ is a bilinear mapping and F_1, F_2 are multiplicative groups, q is a prime and $|F_1| = |F_2| = q$, a is a generator of F_1, $H : F_2 \to \{0,1\}^{\lfloor \log_2(q) \rfloor}$ is a hash function, $x \in G$ is selected randomly.

(2) The encryption E is defined as

$$E_{sk}(m, r) = (c_1, c_2) = H(\hat{e}(x^r, a) \oplus m, a^r),$$

where $m \in \{0,1\}^{\lfloor \log_2(q) \rfloor}$ is a message, $r \in Z_q^*$ is a random value.

(3) The decryption D is defined as

$$D_{sk}(c_1, c_2) = m = H(\hat{e}(c_2, x)) \oplus c_1,$$

this scheme was proved to be IND-CPA secure. The obfuscation process is as follows:

$\underline{\mathcal{O}(E, sk)}$:

— SK is parsed as $(\hat{e},\ F_1,\ F_2,\ q,\ H,\ a,\ x)$;

— Compute $y = \hat{e}(x, a)$;

— Define $pk = (\hat{e},\ F_1,\ F_2,\ q,\ H,\ a,\ x)$;

— Output the encryption $\acute{E}$ as follows: $\acute{E}_{pk}(m, r) = [h(y^r) \oplus m, a^r]$.

In fact, $\acute{E}$ is an asymmetric encryption scheme, it was also proven to be IND-CPA secure. It is to say that the obfuscator keeps the security of original scheme. Therefore, it was claimed that the obfuscation $\mathcal{O}(E, sk)$ satisfies WBP for IND-CPA:

Theorem 3 (Theorem 3 in [22])**.** For a $M \in \mathbb{PTMF} \backslash \mathbb{ALF}$, there is an obfuscator $\mathcal{O} \in \mathbb{PPT}$ and a sn $\in \mathbb{SN}$, s.t. $M \in_{\text{obf}}$ sn and under reasonable computational assumptions, $\mathcal{O}$ satisfies WBP w.r.t (M, sn).

We are not going to repeat the proof here since it is irrelevant to our discussion, for detailed information, refer to Section 5 in [22].

5.2 The overlooked fact in the positive result

In this section, we point out that the security notion "IND-CPA" cannot be used with WBP, so it is not likely that the obfuscator $\mathcal{O}$ satisfies WBP for (E, sn="IND-CPA") in the positive result. Specifically, we prove the following two propositions.

Proposition 1. Security notions defined in black-box attack model cannot be used together with WBP as new security notions.

Proof. We prove our proposition by indicating that the adversaries have different capability in the two kinds of security notions.

First, security notions defined in black-box attack model have a common characteristic: the adversary has only access to the input/output of a cryptosystem (oracle access to the cryptosystem). Without loss of generality, we take security notion "IND-CPA" to show the proof.

It is well known that the definition of IND-CPA is as follows:

Definition 7 (IND-CPA)**.** $A = (A_1, A_2)$ is denoted as an adversary, an asymmetric encryption system $\Pi = (G, E, D)$ is IND-CPA secure if

$$\Pr[(pk, sk) \leftarrow (1^l); (m_0, m_1) \leftarrow A_1^E(pk); b \leftarrow \{0,1\}; y \leftarrow E(m_b) : A_2^E(m_0, m_1, y) = b] \leqslant 1/2 + \mathrm{negl}(l),$$

where A_i^E denotes the adversary A_i can only make oracle queries to E.

However, WBP is defined as

$$|\max(\mathsf{AdvWB}^{\mathrm{sn}}_{A,\mathcal{O},M}(l)) - \max(\mathsf{AdvBB}^{\mathrm{sn}}_{A}(l))| \leqslant \mathrm{negl}(|l|),$$

where $(\mathsf{AdvWB}^{\mathrm{sn}}_{A,\mathcal{O},M}(l))$ is defined with WBgame.

In the WBgame, the adversary has full control of obfuscation $\mathcal{O}(M_i, k_i)$ (e.g. the asymmetric encryption $\acute{E}$ presented in positive result) instead of oracle-access to it, which conflicts with the definition of IND-CPA.

In other words, the adversary in WBP has different capability with the adversary in IND-CPA, thus WBP and IND-CPA cannot be used together as a new security notion, like "WBP w.r.t IND-CPA". This is also suitable to other black-box security notion, such as CCA and CCA-2.

Trivially, Proposition 1 can be extended to the following proposition:

Proposition 2. Without any modification, security notions defined in black-box attack model cannot be used to describe the security of white-box cryptosystem.

6 Conclusion

White-box cryptography (WBC) is becoming popular in real-world applications, which provides opportunities to resist stronger attacks. Many white-box cryptographic schemes have been proposed and analyzed in succession, but only a small amount of theoretical work was published. The results achieved in [22] showed us a theoretical abstraction of white-box security and presented general principles of how to satisfy such security. They have great significance for WBC development.

In this paper, we make discussion on the results in [22]. We showed that the negative result was caused by using an obfuscator which does not satisfy the "Virtual Black-box Property" with auxiliary input, and the positive result cannot hold because the security notion IND-CPA and WBP are not defined for the same adversary.

It should be noted that our discussion aims at accurate understanding of the two results. There still remain some unsolved issues: for negative result, how secure is enough for an obfuscator to instantiate an ideal white-box scheme? For positive result, what security notion can be used in white-box attack context? These issues are exactly our next research work. Meanwhile, we are glad to leave them as open questions and hope they can be solved in the near future.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61272440, 61472251, U1536101), China Postdoctoral Science Foundation (Grant Nos. 2013M531174, 2014T70417) and Science and Technology on Communication Security Laboratory.

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Borghoff J, Canteaut A, Gneysu T, et al. Prince–a low-latency block cipher for pervasive computing applications. In: Advances in Cryptology–ASIACRYPT. Berlin: Springer, 2012. 49–58

2 Wang S B, Zhu Y, Ma D, et al. Lattice-based key exchange on small integer solution problem. Sci China Inf Sci, 2014, 57: 112111

3 Chen Z X. Trace representation and linear complexity of binary sequences derived from Fermat quotients. Sci China Inf Sci, 2014, 57: 112109

4 Chow S, Eisen P, Johnson H, et al. White-box cryptography and an AES implementation. In: Selected Areas in Cryptography. Berlin: Springer, 2003. 250–270

5 Chow S, Eisen P, Johnson H, et al. A white-box DES implementation for DRM applications. In: Digital Rights Management. Berlin: Springer, 2003. 1–15

6 Xiao Y Y, Lai X J. A secure implementation of white-box AES. In: Proceedings of the 2nd International Conference on Computer Science and its Applications, Jeju, 2009. 1–6

7 Karroumi M. Protecting white-box AES with dual ciphers. In: Information Security and Cryptology-ICISC. Berlin: Springer, 2011. 278–291

8 Bringer J, Chabanne H, Dottax E. White box cryptography: another attempt. IACR Cryptology ePrint Archive, 2006, 2011: 468

9 Xiao Y Y, Lai X J. White-box cryptography and a white-box implementation of the SMS4 algorithm. In: ChinaCrypt, Guangzhou, 2009. 24–34

10 Shi Y, Wei W, He Z. A lightweight white-box symmetric encryption algorithm against node capture for WSNs. Sensors, 2015, 15: 11928–11952

11 Link H E, Neumann W D. Clarifying obfuscation: improving the security of white-box DES. In: Proceedings of IEEE International Conference on Information Technology: Coding and Computing, Las Vegas, 2005, 1: 679–684

12 Wyseur B, Michiels W, Gorissen P, et al. Cryptanalysis of white-box DES implementations with arbitrary external encodings. In: Selected Areas in Cryptography. Berlin: Springer, 2007. 264–277

13 Goubin L, Masereel J M, Quisquater M. Cryptanalysis of white box DES implementations. In: Selected Areas in Cryptography. Berlin: Springer, 2007. 278–295

14 Billet O, Gilbert H, Ech-Chatbi C. Cryptanalysis of a white box AES implementation. In: Selected Areas in Cryptography. Berlin: Springer, 2005. 227–240

15 Michiels W, Gorissen P, Hollmann H D L. Cryptanalysis of a generic class of white-box implementations. In: Selected Areas in Cryptography. Berlin: Springer, 2009. 414–428

16 De Mulder Y, Roelse P, Preneel B. Cryptanalysis of the Xiao-Lai white-box AES Implementation. In: Selected Areas in Cryptography. Berlin: Springer, 2013. 34–49

17 Lepoint T, Rivain M, De Mulder Y, et al. Two attacks on a white-box AES implementation. In: Selected Areas in Cryptography–SAC 2013. Berlin: Springer, 2014. 265–285

18 De Mulder Y, Wyseur B, Preneel B. Cryptanalysis of a perturbated white-box AES implementation. In: Progress in Cryptology-INDOCRYPT. Berlin: Springer, 2010. 292–310

19 Lin T T, Lai X J. Efficient attack to white-box SMS4 implementation. J Softw, 2013, 24: 2238–2249

20 Gilbert H, Plt J, Treger J. Key-recovery attack on the ASASA cryptosystem with expanding S-boxes. In: Advances in Cryptology–CRYPTO 2015. Berlin: Springer, 2015. 475–490

21 Herzberg A, Shulman H, Saxena A, et al. Towards a theory of white-box security. In: Emerging Challenges for Security, Privacy and Trust. Berlin: Springer, 2009. 342–352

22 Saxena A, Wyseur B, Preneel B. Towards security notions for white-box cryptography. In: Information Security. Berlin: Springer, 2009. 49–58

23 Saxena A, Wyseur B, Preneel B. White-box cryptography: formal notions and (im) possibility results. IACR Cryptology ePrint Archive, 2008, 2008: 273

24 Valiant L G. A theory of the learnable. Commun ACM, 1984, 27: 1134–1142

25 Linial N, Mansour Y, Nisan N. Constant depth circuits, fourier transform, and learnability. J ACM (JACM), 1993, 40: 607–620

26 Lynn B, Prabhakaran M, Sahai A. Positive results and techniques for obfuscation. In: Advances in Cryptology-EUROCRYPT. Berlin: Springer, 2004. 20–39

27 Wee H. On obfuscating point functions. In: Proceedings of the 37th Annual ACM Symposium on Theory of Computing. New York: ACM, 2005. 523–532

28 Hada S. Zero-knowledge and code obfuscation. In: Advances in Cryptology A SIACRYPT. Berlin: Springer, 2000. 443–457

29 Barak B, Goldreich O, Impagliazzo R, et al. On the (im) possibility of obfuscating programs. In: Advances in cryptology CRYPTO 2001. Berlin: Springer, 2001. 1–18

30 Canetti R, Dakdouk R R. Extractable perfectly one-way functions. In: Automata, Languages and Programming. Berlin: Springer, 2008. 449–460

31 Canetti R, Rothblum G N, Varia M. Obfuscation of hyperplane membership. In: Theory of Cryptography. Berlin: Springer, 2010, 10: 72–89

32 Barak B, Bitansky N, Canetti R, et al. Obfuscation for evasive functions. In: Theory of Cryptography. Berlin: Springer, 2014. 26–51

33 Goldwasser S, Kalai Y T. On the impossibility of obfuscation with auxiliary input. In: Proceedings of IEEE 46th

Annual Symposium on Foundations of Computer Science, Los Alamitos, 2005. 553–562

34 Garg S, Gentry C, Halevi S, et al. Candidate indistinguishability obfuscation and functional encryption for all circuits. In: Proceedings of IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, 2013. 40–49

35 Sahai A, Waters B. How to use indistinguishability obfuscation: deniable encryption, and more. In: Proceedings of the 46th Annual ACM Symposium on Theory of Computing. New York: ACM, 2014. 475–484

36 Hohenberger S, Sahai A, Waters B. Replacing a random oracle: full domain hash from indistinguishability obfuscation. In: Advances in Cryptology-EUROCRYPT. Berlin: Springer, 2014. 201–220

37 Pandey O, Prabhakaran M, Sahai A. Obfuscation-based non-black-box simulation and four message concurrent zero knowledge for np. In: Theory of Cryptography. Berlin: Springer, 2015. 638–667

38 Goldwasser S, Rothblum G N. On best-possible obfuscation. In: Theory of Cryptography. Berlin: Springer, 2007. 194–213

39 Barak B, Goldreich O, Impagliazzo R, et al. On the (im) possibility of obfuscating programs. J ACM (JACM), 2012, 59: 6

40 Bitansky N, Canetti R, Cohn H, et al. The impossibility of obfuscation with auxiliary input or a universal simulator. In: Advances in Cryptology CRYPTO. Berlin: Springer, 2014. 71–89

41 Ananth P, Boneh D, Garg S, et al. Differing-inputs obfuscation and applications. IACR Cryptology ePrint Archive, 2013, 2013: 689

42 Boyle E, Chung K M, Pass R. On extractability obfuscation. In: Theory of Cryptography. Berlin: Springer, 2014. 52–73