SBoxScope: A Meta S-box Strength Evaluation Framework for Heterogeneous Confusion Boxes

Abrar Ahmad, Muddassar Farooq, Amin
Next Generation Intellegent Networks Research Center (nexGIN RC)
Institute of Space Technology
Islamabad, 44000, Pakistan
Email: {abrar.ahmad,muddassar.farooq,amin}@ist.edu.pk

Abstract

In cipher algorithms – both block or streaming – the most important non-linear component is a confusion box (commonly termed as s Substitution box or an S-box). The designers of cipher algorithms create an S-box on the basis of a unique formal model; as a result, its parameters – including its size – are different. Consequently, it becomes a daunting task for a cryptanalyst to conduct a comparative study to analyze, in a scientific yet unbiased manner, the cryptographic strength of these heterogeneous S-boxes. The major contribution of this paper is SBoxScope – a meta S-Box strength evaluation framework – that enables designers and analysts to evaluate cryptographic strength of heterogeneous S-boxes. The framework consists of two layers: (1) White Box Layer analyzes the contents of an S-box and calculates relevant parameters (5 core and 3 auxiliary) and then normalizes them to draw conclusions about the strength of an S-box; (2) Black Box Layer assumes that no knowledge is available about the contents of an S-box; rather, it gives a predefined input bit stream to each S-box and then applies NIST tests to measure parameters. Finally, the two layers are augmented that empowers an analyst to make a decision about the strength of an S-box after analyzing different parameters. In this paper, we have evaluated 9 S-boxes of five well known cipher algorithms: AES, MARS, Skipjack, Serpent and Twofish.

Keywords

Cryptography, Cipher Algorithms, Cryptographic Strength

1. Introduction

The majority of cipher algorithms, proposed in the literature, uses a non-linear component – a Substitution box (S-box) – that introduces confusion in a cipher. The motivation of using a cryptographically strong S-box is to make it difficult for a cryptanalyst to break a cipher by modeling it with the help of well known linear functions. The designers of cipher algorithms create an S-box with the help of highly non-linear formal models. As a result, every S-box is unique having its own set of parameters: (1) size of S-box (number of elements); (2) the number of rows and columns of an S-box; (3) the size of substitution block (nibble, byte or 32 bits word). An important challenge for a cryptanalyst is: how to conduct a comparative study to draw scientific conclusions about the cryptographic strength of heterogeneous S-boxes? To the best of our knowledge, no framework exists that empowers a designer or cryptanalyst to determine with a certain degree of confidence e.g. the S-box of AES is cryptographically stronger compared with that of MARS.

The major contribution of this paper is a meta S-box strength evaluation framework - SBoxScope – that allows S-box designers to compare the cryptographic strength of heterogeneous S-boxes. SBoxScope consists of two layers: White Box Layer (WBL) and Black Box Layer (BBL). The WBL assumes that the design, implementation, and contents of an S-box are known and it conducts a pattern based analysis on the contents by running 8 independent tests. Consequently, 5 core and 3 auxiliary parameters are computed by the WBL. These parameters are then normalized to remove the bias introduced due to the size of an S-box. Finally, it ranks the S-boxes on each parameter and accumulates the rank score of each S-box. An S-box with the lowest rank score is declared as the best cryptographically strong S-box among all heterogeneous S-boxes.

The BBL assumes that a cipher algorithm consists of only one component – the S-box – and its design, implementation and contents need not to be known. BBL creates a number of input bit streams by taking inspiration from NIST test suite for cipher algorithms. It provides these streams to each S-box and its output is treated as the cipher text. In the next step 10 NIST tests are conducted and the p-values of relevant parameters are measured. The BBL provides the analysts a microlevel probe into the substitution pattern of an S-box. The analysts could easily compare the p-values of input bit stream with that of the output cipher stream. BBL provides a valuable insight into an S-box: if an S-box is the only component of a cipher algorithm then how much
TABLE 1: Design Parameters of Shortlisted S-boxes

<table>
<thead>
<tr>
<th>SBox</th>
<th>Algorithm</th>
<th># of elements</th>
<th>Dimension</th>
<th>Bits Substituted</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_A$</td>
<td>AES</td>
<td>256</td>
<td>16 x 16</td>
<td>8</td>
</tr>
<tr>
<td>$S_M$</td>
<td>MARS</td>
<td>512</td>
<td>512 x 1</td>
<td>32</td>
</tr>
<tr>
<td>$S_K$</td>
<td>Skipjack</td>
<td>256</td>
<td>16 x 16</td>
<td>8</td>
</tr>
<tr>
<td>$S_0$</td>
<td>Serpent</td>
<td>16</td>
<td>1 x 16</td>
<td>4</td>
</tr>
<tr>
<td>$S_4$</td>
<td>Serpent</td>
<td>16</td>
<td>1 x 16</td>
<td>4</td>
</tr>
<tr>
<td>$S_7$</td>
<td>Serpent</td>
<td>16</td>
<td>1 x 16</td>
<td>4</td>
</tr>
<tr>
<td>$S_{T1}$</td>
<td>Twofish</td>
<td>16</td>
<td>1 x 16</td>
<td>4</td>
</tr>
<tr>
<td>$S_{T2}$</td>
<td>Twofish</td>
<td>16</td>
<td>1 x 16</td>
<td>4</td>
</tr>
<tr>
<td>$S_{T3}$</td>
<td>Twofish</td>
<td>16</td>
<td>1 x 16</td>
<td>4</td>
</tr>
</tbody>
</table>

confusion it can add, without the need of complementing it with a Substitution Permutation Network (SPN), to make it infeasible for a cryptanalyst to draw conclusions about the correlation between ciphertext and plaintext. Finally, BBL ranks each S-box on the basis of p-value of each parameter and then accumulates the rank score. Similar to WBL, an S-box with the lowest rank score is considered to be the best cryptographically strong S-box.

Once the S-boxes are ranked by WBL and BBL, SBoxScope classifies only those S-boxes as cryptographically strong that are ranked high (closest to their ideal values) both by WBL and BBL; if an S-box does not achieve a relatively high rank in any of the testing layers, it is dropped from the list of cryptographically strong S-boxes.

In order to do the comparative study, we surveyed more than 30 well known cipher algorithms and shortlisted only 5 among them for the study. The five algorithms are: AES [11], MARS [12], Skipjack [15], Serpent [13] and Twofish [14]. The 4 shortlisted algorithms (except Skipjack) reached the final of NIST competition held for standardizing AES. (We have excluded RC6 because it does not have an S-box). Skipjack is shortlisted because it has been widely used by NSA and the structure of its S-box (16 * 16 elements) is quite similar to that of AES. (NSA declassified Skipjack on 29th May 1998) [17]. AES, MARS and Skipjack have only 1 S-box each. Twofish and Serpent have similar kind of 8 S-boxes. In case of Twofish, we have chosen two S-boxes from q0 permutation and one S-box from q1 permutation steps. Similarly, we have chosen s0, s4 and s7 S-boxes for Serpent. As a result, we have shortlisted 9 S-boxes of the above-mentioned five algorithms for our comparative study.

Table 1 tabulates different design parameters of these S-boxes that demonstrates diversity and heterogeneity of shortlisted S-boxes. The results of SBoxScope show that the S-boxes of AES and MARS have consistently outperformed other S-boxes on the majority of WBL and BBL strength parameters and clearly classified as the two best cryptographically secure S-boxes among the selected 9 candidates. The Skipjack S-box is placed at number 3 after AES and MARS S-boxes.

Organized of the Paper: The rest of the paper is organized as follows. In Section 2, a brief summary of the work that is relevant and related to SBoxScope is presented. Section 3 discusses the architecture of SBoxScope and also gives definitions of quantitative parameters used in WBL and BBL. The section also builds an intuitive understanding of each parameter by highlighting what aspect of strength it measures. After presenting the quantitative framework of SBoxScope, the detailed discussion and comprehensive analysis of the results obtained from the SBoxScope are presented in Section 4. Finally, we conclude the paper with an outlook to our future work in Section 5.

2. Related Work

To the best of our knowledge, no framework is presented in the literature that allows analysts to compare cryptographic strength of heterogeneous S-boxes. Having said that, it is important to emphasize that the core parameters that build the foundation of quantitative framework of WBL and BBL are already present in the literature. The novel contribution is to adapt and normalize them for an unbiased comparison of different heterogeneous S-boxes. For example, WBL uses five core parameters – Nonlinearity (NL) [6], Bit Independence Criteria (BIC) [4], Strict Avalanche Criteria (SAC) [4], Differential Probability (DP) and Linear Probability (LP) [7][8] – to measure the cryptographic strength of S-boxes. The NL is defined by Meier and Staffelbach in [6]. Similarly BIC and SAC were introduced by Webster and Tavares in [4]. Generally speaking, these parameters are correlated: an S-box with high nonlinearity will also have high BIC and nearly desirable SAC (approximately 0.5).

The WBL of SBoxScope consists of 3 auxiliary parameters – Fixed Points (FP), Balancedness (BN) and Correlation Immunity (CI) – that are referred to if the ranks computed on the basis of core parameters, of two or more S-boxes are very close to each other. FP, BN and CI are proposed in [10], [9] and [6] respectively. The desirable value for BN and FP is zero; while for CI it is 1.

The BBL of SBoxScope is adapted from NIST test suite [1] that is the standard benchmark for evaluating cipher algorithms. The test suite consists of generating 8 different types of input bit streams and then running 15 tests on the cipher streams and computing associated 41 parameters for each type of cipher stream. NIST recommends computing p-values of each parameter to conclude whether the cipher algorithm has passed the corresponding test or not. After carefully analyzing the NIST test suite, we have shortlisted only two bit stream generation methods that are relevant to the context of S-box evaluation: Cipher Text Plain Text Correlation (CTPT) and Random Plain Text (RPT). The other 6 NIST stream generation methods are relevant for evaluating entire cipher algorithms and it is beyond the scope of this paper. Similarly, instead of running 15 tests, BBL only conducts 10 tests: Frequency Test, Block Frequency Test, Run Test, Long Run Test, Rank Test, Discrete Fourier Transform Test, Linear Complexity Test, Approximate Entropy Test, Random...
Excursion Test and Random Excursion Variate Test (Refer to [1] for details). The remaining 5 tests – Non-Overlapping Test, Overlapping Test, Universal Statistical Test, Serial Test and Cumulative Sum Test – are relevant for evaluating an entire cipher algorithm and hence are excluded from BBL of SBoxScope. The idea of adapting and applying NIST test suite for Black Box testing of S-boxes is novel and helps in answering a fundamental question: If a given S-box is not complemented by an SPN network, then how much confusion it can add to the substituted cipher stream? The answer provides useful insight into the substitution pattern of a given S-box and its associated strength.

3. The Architecture of SBoxScope

The architecture of SBoxScope is presented in Figure 1. The SBoxScope consists of four major components: (1) Adaptation Layer; (2) White Box Layer; (3) Black Box Layer and (4) Ranking Layer. The adaptation layer converts the contents of S-boxes into a byte stream and also takes its Walsh transform. Walsh Transform is basically representation of Boolean functions with values +1 and -1 that is used to compute Non-linearity [6]. Both S-box and its Walsh transform are given as input to the WBL. Moreover, it also adapts the core and auxiliary parameters to enable WBL to compute these for any type of S-box. As mentioned before, WBL computes 5 core and 3 auxiliary parameters for each S-box. The values of parameters are stored in a database. The ranking layer is then utilized to rank S-boxes for each parameter. If the rank of two or more S-boxes, determined by core parameters, is close to each other then WBL considers their rank on the basis of auxiliary parameters.

The Black Box Layer (BBL) builds a .dll of each S-box by treating an S-box as the cipher algorithm. BBL then applies adapted NIST test suite on each S-box and stores p-values of associated parameters in a database. Since, NIST recommends an $\alpha = 0.01$; therefore, if $p$-value $\geq 0.01$, it means the sequence after substitution would be considered to be random with a confidence level of 99% [1]. The ranking layer again ranks each S-box on the basis of p-values for each test and then determines the overall rank of an S-box on the basis of all 10 tests. (All symbols used in this paper are tabulated in Table 2 for a ready reference.)

Quantitative Measurement Framework for White Box and Black Box Layers:

WBL analyzes the implementation of an S-box by treating it as a substitution table of a certain size and analyzing the values in the S-box. We now describe each core and auxiliary parameter, computed by WBL layer, and intuitively build its relevance to the cryptographic strength of a given S-box.

3.1. White Box Layer (WBL) Parameters

Nonlinearity. The Nonlinearity $N_f$ of a boolean function is defined as the distance between the function and Affine
functions that are composition of linear functions and their translations [6]. Mathematically, it is defined by:

\[ N_f = 2^{n-1} - \frac{1}{2}\max\{|\langle\lambda, l_i\rangle|, 0 \leq i \leq 2^n - 1\}, \]  

where \( \lambda \) is the sequences of functions and \( l_0, \ldots, l_{2^n-1} \) are rows of Hadamard Matrix used in calculating nonlinearity.

Generally speaking, a higher value of nonlinearity is desirable because it shows that it is not possible to estimate the substitution pattern of an S-box with the help of Affine functions. The other challenge is: the simple value of \( N_f \) cannot be used to quantitatively compare different heterogeneous S-boxes because in each case the ideal value of nonlinearity is different and depends on its size. For example the ideal value of 256 elements of AES S-box is 120; while for 512 element of MARS S-box, it is 210. In order to remove the bias of size, WBL normalizes the nonlinearity in the following manner:

\[ \eta_N = \frac{\log_2(\text{obtained } N_f)}{\log_2(\text{ideal } N_f)} \]  

The upper bound for nonlinearity (or ideal nonlinearity) is computed by using the formula given in [6]. The range of normalized nonlinearity is \( 0 < \eta_N \leq 1 \) that is independent of the size of an S-box.

**BIC** defines that all avalanche variables should be pairwise independent for a given set of avalanche vectors that are generated by complementing a single plain text bit[16]. It ensures that an S-box with one changed bit could still not be modeled by Affine functions. BIC also has an ideal upper value for each type of S-box and WBL normalizes its value by using the above-mentioned normalization formula. Similarly, the range of BIC is \( 0 < \beta_N \leq 1 \).

**SAC** is represented by \( \sigma \) that ensures if one bit is changed at the input, approximately half the bits are changed at the output. Its ideal value is 0.5 and it is independent of the size of S-box which does not require normalization.

**Differential Probability (DP)** ensures that an input differential \( \Delta x_i \) shall uniquely map to an output differential \( \Delta y_i \) [16]. Consequently, if any number of bits are changed at the input, a constant or uniform number of bits are always changed at the output and this is termed as Uniform differential criterion. In order to compute DP, one has to take a fraction of the uniform number of bits changed to the total number of elements in the S-box. For AES S-box, DP is \( \frac{2}{256} = 0.0156 \). Similarly, for the S-box of Twofish it is \( \frac{2}{256} = 0.125 \). Since DP is dependent on the size of S-box; therefore, WBL normalizes it using the following formula:

\[ \eta_D = \frac{\text{(obtained } DA)\text{}}{n}, \]  

where \( n \) is number of bits substituted by an S-box and Differential Approximation (DA) is defined by [7][8]:

\[ DA = \#\{x \in X/S(x) \oplus S(x \oplus \Delta x) = \Delta y\}, \]  

DA defines the number of times an S-box fulfills uniform differential criterion. Ideally \( \eta_D \) should be 0.5. After doing the normalization, it is possible to compute and compare DP of any type of S-box.

**Linear Probability** is the maximum value of the imbalance of an event. The parity of the input bits selected by the mask \( \Gamma_x \) is equal to the parity of the output bits selected by the mask \( \Gamma_y \) [16]. LP is defined in [7][8] as:

\[ \eta_L = \max_{\Gamma_x, \Gamma_y \neq 0} \left| \frac{\#\{x|x \cdot \Gamma_x = S(x) \cdot \Gamma_y\}}{2^n} - \frac{1}{2} \right| \]  

It ensures that by knowing the cipher stream, mask \( \Gamma_x \) and mask \( \Gamma_y \), an analyst should not be able to predict the plain text. The ideal value of LP is 0.

**Fixed Point (FP)** of an S-box is defined by:

\[ f(x) = x, \]  

This means that the number of times the output (after substitution) is the same as input. WBL normalizes Fixed Points (FP) as:

\[ \xi = \frac{\text{(obtained } \# \text{ of FP})}{2^n} \]  

Its ideal value is zero.

**Balancedness (BN)** A boolean function \( f : \{0, 1\}^n \rightarrow \{0, 1\} \) is said to be balanced if its truth table has \( 2^{n-1} \) zeros (or ones) [9]:

\[ \sum_{w \in \{0, 1\}^n} f(w) = 2^{n-1} \]
In order to compute BN, WBL converts all entries of an S-box into binary and writes them into a table. If the number of zeros and ones in each of the 8 bit columns is the same, the corresponding S-box is considered as balanced and BN = 0; otherwise, BN is equal to the number of unbalanced columns. The BN is normalized using the following equation:

\[ \beta_n = \frac{\text{obtained BN}}{n}, \]  

(9)

where \( n \) is the number of bits substituted by an S-box.

**Correlation Immunity (CI)** ensures that the elements of an S-box should be independent of each other; as a result, the substituted cipher stream does not contain correlated sequences. WBL normalizes the CI by using the following equation:

\[ \psi = \log_2(\text{obtained CI}) - 1 \]

(10)

Ideal value of \( \psi \) is 1.

### 3.2. Black Box Layer (BBL) Parameters

As mentioned before, we have used only 2 bit stream generation methods that are relevant to evaluate the strength of S-boxes: (1) Cipher Text Plain Text (CTPT) correlation (recommended by NIST) in which pseudo-random number generator (PRNG) is applied at input (see Fig 2); (2) Random Plain Text (RPT), running a stress test on S-Boxes by generating input stream from a highly correlated windows image (see Fig 4). In CTPT, the input generated stream is given to the S-box to get the substituted cipher stream. Afterwards, both streams are exclusively xored and NIST tests are run on the final stream. The basic motivation of CTPT is to factor out the correlative artifacts in the plain bit stream; and instead focus on the randomness in the substituted cipher bit stream. Once a test is finished, the \( p \)-value of the associated parameter is computed. If the \( p \)-value \( \geq 0.01 \), S-box passes the test; else fails. Now the purpose of each test is briefly summarized as described by the authors in [1] to make the paper self contained. An interested reader may find detailed discussion in [1].

**Frequency Test (F).** The purpose of this test is to determine whether an S-box is able to ensure that the number of ones and zeros in the substituted cipher stream are approximately the same as would be expected in a random cipher. The \( p \)-value of the associated parameter is \( P_F \).

**Block Frequency Test (BF).** Ensures that the S-box is able to maintain the notion of randomness – equal number of ones and zeros – even in small substituted blocks of a given length \( M \). Its \( p \)-value is denoted by \( P_B \).

**Runs Test (Rn).** The purpose of this test is to determine whether the S-box is able to maintain the required oscillation speed between variable length \( k \) continuous ones and zeros. The test identifies whether the transitions between such zeros or ones is too slow or too fast. Its \( p \)-value is denoted by \( P_R \).

**Longest Run of Ones in a Block Test (LR).** The purpose of this test is to determine whether the S-box is able to limit the longest run of ones within \( M \) block bits in such a fashion as expected in a random bit stream. Consequently, if the longest run of ones is irregular, the same would hold for zeros. Its value is denoted by \( P_L \).

**Binary Matrix Rank Test (RK).** The purpose of this test is to ensure that S-box should not introduce a linear dependence among fixed length disjoint sub matrices of the entire cipher bit stream. Its \( p \)-value is denoted by \( P_K \).

**Discrete Fourier Transform Test (DFT).** The purpose of this test is to identify whether the S-box has introduced periodic features in the cipher bit stream that would indicate a deviation from assumed randomness. The intention is to detect whether the number of peaks, in the Discrete Fourier Transform of cipher bit stream, exceeding the 95% threshold differs significantly by 5%. Its \( p \)-value is denoted by \( P_D \).

**Linear Complexity Test (LC).** The purpose of this test is to determine randomness, introduced by the S-box, in the cipher stream by computing the length of Linear Feedback Shift Register. Longer LFSR characterizes a random sequence. Its \( p \)-value is denoted by \( P_C \).
Approximate Entropy Test (AE). The purpose of this test is to determine whether an S-box has introduced overlapping m-bits patterns in the substituted cipher stream. A large frequency of consecutive m and m+1 length blocks is deviation from randomness. Its p-value is denoted by \( P_A \).

Random Excursion Test (RE). The purpose of this test is to determine if the number of visits to a particular state within a cycle – consisting of a sequence of steps of unit length taken at random in such a fashion that one returns to origin – deviates from what one would expect for a random sequence. In this test, (0,1) is transformed to (-1, +1) and then the number of visits to -4, -3, -2, -1, and +1, +2, +3 and +4 are calculated; as a result, we get 8 p-values corresponding to each state. To simplify analysis, BBL selects the minimum among them. It is denoted by \( P_E \).

Random Excursion Variant Test (REV). The purpose of this test is to determine the number of times a particular state is visited in cumulative sum random walk and then conclude whether it deviates from the random walk. This test consists of a series of 18 tests and produces 18 p-values. BBL again picks up the minimum one among them to simplify the analysis. The p-value is denoted by \( P_V \).

If an S-box has passed the BBL test, then we classify it into one of 5 classes (see Table 4) depending on the p-value.

4. Discussion and Results

We now present the results obtained, with the help of SBoxScope, from the comprehensive experiments. The experiments were conducted on a Virtual Machine (VM), obtained from an HP Server running VMware. The specifications of the VM are: (1) QuadCore 2 GHz processor; (2) 8 GB RAM; and (3) 40 GB Hard Disk. SBoxScope is implemented in C# as a multi-threaded application. The powerful VM allows to create and execute 8 threads in parallel; as a result, the WBL parameters are computed in less than 2 minutes and stored in an oracle database. Moreover, BBL parameters for one input stream – CTPT or RPT – and one S-box are computed in less than 10 hours; as a consequence, all nine S-boxes for one input stream are evaluated in less than 4 days.

We have shortlisted 9 S-boxes of 5 well known algorithms: AES, MARS, Skipjack, Serpent and Twofish. WBL computes 5 core parameters and 3 auxiliary parameters; while BBL computes 10 p-values for 10 shortlisted tests. In our discussion on WBL results, we will particularly try to answer the following questions:

1) What is the relationship of normalized Nonlinearity with the strength of an S-box?
2) Whether the generally perceived fact holds for the selected S-boxes that a highly nonlinear S-Box will also have a large BIC and nearly ideal SAC value?
3) What conclusions about the strength of an S-box can be drawn from DP and LP?
4) If some S-boxes cannot be differentiated on the basis of core parameters, could they be differentiated on the basis of Auxiliary parameters?

Similarly, for BBL results, we will address the following pertinent issues:

1) What is the role of input stream generation method on p-values of shortlisted BBL tests?
2) What is the role of an S-box on p-values of shortlisted BBL tests?
3) Whether the hypothesis holds that a strong S-box, declared by WBL, is expected to pass all BBL tests?
4) How significant it is that a strong S-box, declared by WBL, fails a BBL test?
5) Is it possible to generalize with a reasonable degree of accuracy that a strong S-box, declared by WBL, will pass BBL tests with higher p-values?
6) Whether SBoxScope is able to differentiate strong S-boxes from weak ones by comprehensively looking at WBL and BBL decisions?

4.1. Results and Discussions of White Box Layer (WBL) Tests

The 5 core and 3 auxiliary parameters, computed by WBL, are enumerated in Table 3. Now let us try to answer the questions related to WBL in the previous Section.

(1) What is the relationship of normalized Nonlinearity with the strength of an S-box?

It is obvious from the nonlinearity results in Table 3 that AES S-box has the highest normalized nonlinearity (\( \eta_N \)) followed by MARS and Skipjack S-boxes. This shows that these S-boxes have been able to achieve approximately their respective ideal nonlinearity values. From the results, we can easily make a hypothesis: One of these 3 S-boxes is expected, with a high probability, to be declared as the strongest cryptographic S-box. In comparison, Serpent and Twofish S-boxes have very small nonlinearity values and therefore may not be able to do well in the SBoxScope investigations.

(2) Whether the generally perceived fact holds for the selected S-boxes that a highly non-linear S-Box will also have a large BIC and nearly ideal SAC value?

If we look at the normalized BIC (\( \eta_{ic} \)) and SAC (\( \sigma \)) values in Table 3, it is easy to assert that the hypothesis holds: highly nonlinear S-boxes (that of AES, MARS and Skipjack) also have nearly ideal BIC and SAC values.

(3) What conclusions about the strength of an S-box can be drawn from DP and LP?

It is interesting to note that all S-boxes have nearly ideal DP (\( \eta_{DP} \)) value – 0.5 – except the MARS S-box that has 0.445. Nevertheless it shows that if any number of bits change at the input, almost all S-boxes are able to make sure that the output changes by a fixed number of bits at least in half the elements
of S-box. To conclude, DP is an important and necessary criterion but it does not provide sufficient information to determine the strength of an S-box.

In case of normalized LP $\eta_L$, it is clear from Table 3 that AES and MARS S-boxes have achieved values closer to the ideal value of 0, whereas for Serpent and Twofish S-boxes, this is significantly lower; therefore, they are unable to maintain a desirable uniform parity bits pattern in the cipher stream. As a result, an analyst can detect parity patterns in the cipher stream. Serpent S-boxes have the worst $\eta_L$ values. To conclude, $\eta_L$ may provide sufficient information to determine the strength of an S-box. If we look at the cumulative rank score, computed by WBL on the basis of core parameters, S-box of AES has outperformed the others followed by that of MARS and Skipjack. However, one cannot make a statistically significant difference between MARS and Skipjack S-boxes and the same holds for Serpent and Twofish S-boxes, on the basis of core parameters.

(4) If some S-boxes cannot be differentiated on the basis of core parameters, could they be differentiated on the basis of Auxiliary parameters?

Now SBoxScope brings into the picture Auxiliary parameters, computed by WBL, to differentiate the closely ranked S-boxes. It is obvious from Table 3 that $S_4$ of Serpent and $S_{T1}$, $S_{T2}$ of Twofish have fixed points which is not a desirable characteristic. Similarly, all S-boxes have normalized BN ($\beta_N$) equal to the ideal value of 0 except MARS and Skipjack. In case of MARS all columns do not have equal number of zeros and ones (though the imbalance is very small); while in case of Skipjack a quarter of the columns are not balanced. Finally, normalized CI ($\psi$) of Skipjack S-box is the highest that is closely followed by AES and MARS. The values for other S-boxes are significantly inferior. The aforementioned discussion clearly suggests that $\xi$ and $\beta_N$ are important requirements but not sufficient to determine the strength of S-boxes; however, $\psi$ may be an additional parameter that distinguishes the strength of an S-box.

The auxiliary parameters hence, may help to further distinguish two closely ranked S-boxes on the basis of $\psi$, this distinction however may not be conclusive for other auxiliary parameters ($\xi, \beta_N$). The final rank of all nine S-boxes is shown in Table 3.

The major conclusion of WBL analysis is: AES, MARS and Skipjack S-boxes are cryptographically strong and it is difficult for a cryptanalyst to launch linear or differential attacks on them. In comparison, Twofish and Serpent S-boxes are not robust against linear cryptanalysis due to small $\eta_N$, large $\eta_L$ and large $\psi$; hence, they must be complemented through a SPN network in a cipher algorithm.

### 4.2. Results and Discussions of Black Box Layer (BBL) Tests

BBL creates two input bit streams: CTPT and RPT. BBL runs 10 tests mentioned in Section 3.2 and then computes p-values. Subsequently, it classifies the S-box on the basis of p-values with the help of classes mentioned in Table 4. Now, we try to answer the questions related to BBL testing mentioned in the previous section.

(1) What is the role of input generation stream method on p-values of shortlisted BBL tests?

It is important to again emphasize that SBoxScope assumes that the S-box is the only component of a cipher algorithm; therefore, it analyzes its strength by running WBL and BBL tests. If we look at Table 5 and Table 6, and compare the p-values of S-boxes with that of the original input stream, then we can easily conclude that in case of CTPT (logical stream generation mechanisms as prescribed by NIST), all S-boxes have passed all 10 tests of BBL except $S_0$ S-box of Serpent that has failed Approximate Entropy test ($P_A$). This may be compared with the original p-values of the Windows Image in Table 6 and they represent a significant challenge for all S-boxes. As expected, all S-boxes have failed Ranks test ($P_K$) and Linear Complexity Test $P_C$ because highly

### TABLE 3: White Box Layer (WBL) Tests: Core and Auxiliary Parameters

<table>
<thead>
<tr>
<th>Sr</th>
<th>SBox</th>
<th>$\eta_N$</th>
<th>$\beta_L$</th>
<th>$\sigma$</th>
<th>$\eta_D$</th>
<th>$\eta_L$</th>
<th>$A$</th>
<th>$\xi$</th>
<th>$\beta_N$</th>
<th>$\psi$</th>
<th>$B$</th>
<th>Final = $A+B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$S_A$</td>
<td>0.985</td>
<td>1</td>
<td>0.062</td>
<td>1</td>
<td>0.500</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1.433</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>$S_M$</td>
<td>0.977</td>
<td>2</td>
<td>0.807</td>
<td>2</td>
<td>0.445</td>
<td>2</td>
<td>0</td>
<td>0.972</td>
<td>2</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>$S_K$</td>
<td>0.952</td>
<td>3</td>
<td>0.974</td>
<td>3</td>
<td>0.500</td>
<td>1</td>
<td>0.113</td>
<td>3</td>
<td>1.263</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>$S_6$</td>
<td>0.774</td>
<td>4</td>
<td>0.774</td>
<td>4</td>
<td>1.000</td>
<td>5</td>
<td>0.500</td>
<td>1</td>
<td>0.250</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>$S_4$</td>
<td>0.774</td>
<td>4</td>
<td>0.774</td>
<td>4</td>
<td>1.000</td>
<td>5</td>
<td>0.500</td>
<td>1</td>
<td>0.250</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>$S_7$</td>
<td>0.774</td>
<td>4</td>
<td>0.774</td>
<td>4</td>
<td>1.000</td>
<td>5</td>
<td>0.500</td>
<td>1</td>
<td>0.250</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>$S_{T1}$</td>
<td>0.386</td>
<td>5</td>
<td>0.774</td>
<td>4</td>
<td>1.000</td>
<td>5</td>
<td>0.500</td>
<td>1</td>
<td>0.375</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>$S_{T2}$</td>
<td>0.774</td>
<td>5</td>
<td>0.774</td>
<td>4</td>
<td>1.000</td>
<td>5</td>
<td>0.500</td>
<td>1</td>
<td>0.250</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>$S_{T3}$</td>
<td>0.774</td>
<td>5</td>
<td>0.774</td>
<td>4</td>
<td>1.000</td>
<td>5</td>
<td>0.500</td>
<td>1</td>
<td>0.375</td>
<td>5</td>
<td>18</td>
</tr>
</tbody>
</table>

### TABLE 4: BBL Classification of S-box on p-values

<table>
<thead>
<tr>
<th>Sr</th>
<th>$\xi$</th>
<th>Class</th>
<th>Class Definition</th>
<th>BBL Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$\xi = 1$</td>
<td>$0.5 \leq p-value$</td>
<td>Extremely Strongly Passed</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$\xi = 2$</td>
<td>$0.4 \leq p-value &lt; 0.5$</td>
<td>Strongly Passed</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>$\xi = 3$</td>
<td>$0.3 \leq p-value &lt; 0.4$</td>
<td>Moderately Passed</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>$\xi = 4$</td>
<td>$0.2 \leq p-value &lt; 0.3$</td>
<td>Satisfactorily Passed</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>$\xi = 5$</td>
<td>$0.01 \leq p-value &lt; 0.2$</td>
<td>Barely Passed</td>
</tr>
<tr>
<td>6</td>
<td>$F$</td>
<td>$p-value &lt; 0.01$</td>
<td>Failed</td>
<td></td>
</tr>
</tbody>
</table>
correlated data exists in the picture. Moreover, Serpent and Twofish have even failed 3 to 4 more tests. To conclude, it is important to generate pseudo-random input stream; otherwise, BBL analysis would not be very useful.

(2) What is the role of an S-box on p-values of shortlisted BBL tests?

Generally speaking it is evident from Table 5 that p-values are dependent on the S-box. A cryptographically strong S-box has passed all tests with high p-values.

(3) Whether the hypothesis holds that a strong S-box, declared by WBL, is expected to pass all BBL tests?

The results in Table 5 and Table 6 prove the hypothesis that cryptographically strong S-boxes, determined by WBL, have passed BBL tests with relatively high p-values. In fact, the final rank score in Table 5 proves that AES, MARS and Skipjack S-boxes are not only the top ranking S-boxes in WBL but also in BBL. Similarly, Serpent and Twofish S-boxes are poorly ranked by both WBL and BBL of SBoxScope.

(4) How significant is it that a strong S-box, declared by WBL, fails a BBL test?

If the p-values of input data stream are in the passing range and then an S-box fails a test, it is a significant weakness. On the other hand, if the p-values of input data stream are small (or in the failing range), then a test failure is undesirable but may not be avoidable due to correlation in the input data.

(5) Is it possible to generalize with a reasonable degree of accuracy that a strong S-box, declared by WBL, will pass BBL tests with higher p-values?

Yes and the discussion is already done in answering question number 2.

(6) Whether SboxScope will be able to differentiate strong S-boxes from weak ones by looking at WBL and BBL decisions?

If we look at the ranking of S-boxes (determined by WBL) in Table 3 and compare them with the ranking of S-boxes (determined by BBL) in Tables 5 and 6, it is easier to assert that SBoxScope has been able to differentiate strong S-boxes of AES, MARS and Skipjack from weak S-boxes of Serpent and Twofish. It is again emphasized that Serpent and Twofish themselves are strong ciphers because they complement their weak S-boxes with a strong SPN. The obvious reason for using small S-boxes appear to be memory and processing efficiency.

To conclude, the overall rank (1 being the highest and 9 being the lowest) of 9 S-boxes, determined by SBoxScope by adding scores of WBL and BBL tests is: (1) AES S-box; (2) MARS S-box; (3) Skipjack; (4) S4 of Serpent; (5) S7 of Serpent; (6) ST2 and ST3 of Twofish; (7) ST3 of Twofish; and (8) S6 of Serpent.
5. Conclusions and Future Work

The major contribution of this paper is SBoxScope that has the ability to determine the cryptographic strength of heterogeneous S-boxes and rank them after removing biasness due to size of the S-box. SBoxScope consists of two layers: (1) White Box Layer assumes that the design of an S-box is known and it computes 5 core and 3 auxiliary parameters; and (2) Black Box Layer assumes that the implementation of S-box is not known and it runs adapted NIST test suite to determine whether an S-box can pass the tests on its own or not. BBL also classifies how strongly an S-box has passed a test.

We evaluated 9 S-boxes of five well known cipher algorithms – AES, MARS, Skipjack, Serpent, and Twofish – with the help of SBoxScope. The outcome of the comprehensive analysis is that AES S-box is at the top of the list among all S-boxes followed by MARS S-box and Skipjack S-box respectively. The other important contribution of the study is that small size S-boxes need to be complemented with a strong SPN. Probably, the reason of choosing small S-boxes appear to be memory and processing efficiency. These might be relevant at the time NIST announced the competition but in the last decade, the cost of high speed processors and memory has significantly reduced. Therefore, it is not a good idea to use large number of relatively small size cryptographically weak S-boxes.

In future, we want to evaluate S-boxes of all 30 surveyed cipher algorithms and also increase the unit of evaluation to a round level to better compare the cryptographic strength of cipher algorithms. This shall be the topic of forthcoming publications.

Acknowledgment

Abrar Ahmad is supported through a research fellowship (Ref no: IST/R&D/30026) awarded by IST R&D Fund. The information, data, comments and views detailed herein may not necessarily reflect the endorsements of IST R&D Fund.

References

[3] Kavut, S. Yucel, M.D, There exist Boolean function on n (odd) variables having nonlinearity $2^n-1-2^{\lfloor (n-1)/2 \rfloor}$ if and only if n $\geq$ 7. IACR Cryptology ePrint Archive 2006, 181 (2006)