Research on the Grouping Method of Side-Channel Leakage Detection

Xiaoyi Duan, Ye Huang, YongHua Su, Yujin Li, and XiaoHong Fan(✉)

Beijing Electronic Science and Technology Institute, Beijing, China
fanxiaohong@139.com

Abstract. Power analysis attack is a method to obtain the key of the cryptographic chip by analyzing the correlation between power consumption information leaked by the cryptographic chip during the computing process and the key. The efficiency of power analysis attack poses a serious threat to the software and hardware implementation of cryptographic algorithms. In order to detect whether a cryptographic chip has information leakage, it is necessary to assess it by using detection techniques. The t-test is a hypothesis test used in the field of statistics to test whether there is a significant difference in the means of two normally distributed populations with unknown variance, and is also a useful tool in side-channel information leakage assessment. In this paper, two grouping methods are proposed based on the characteristics of the AES algorithm to investigate the construction of two overall groups before the implementation of the Welch's t-test. Experimental verification of the DPA contest V4 dataset shows that both grouping methods were effective in detecting a large number of leakage points on power traces, but the grouping method by AES first round S-box output Hamming weight has a higher proportion of both the number of leakage points and the high t-statistic distribution than the method of grouping by bit value.

Keywords: Leakage detection · Welch's T-test · AES

1 Introduction

With the widespread use of cryptographic devices such as smart cards, the security of cryptographic chips has become a major concern as a security safeguard for cryptographic devices. The key determines the security of the cryptographic algorithm, so the attackers often target the key of the cryptographic algorithm. Traditional brute-force attacks using hardware and software are time-consuming and inefficient. In recent years, the emergence of Side Channel Attack (SCA) has allowed information such as power, runtime and electromagnetic radiation leaked during the operation of cryptographic devices to be used by attackers to analyze the correlation with intermediate values of cryptographic algorithms and ultimately to break keys. The operation of the cryptographic algorithm in the cryptographic chip will result in the leakage of a lot of information. The power analysis attack on key information is an important part of the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
F. Li et al. (Eds.): SecureComm 2022, LNICST 462, pp. 807–818, 2023.
https://doi.org/10.1007/978-3-031-25538-0_42

side-channel attack, which attempts to decipher keys by collecting the power consumption information leaked by the cryptographic chip during the operation and analyzing the relationship between the power consumption values and keys [1]. Generally, power analysis attack techniques have been mature, commonly used methods are Simple Power Analysis(SPA), Differential Power Analysis [2](DPA), Correlation Power Analysis [3] (CPA), template attack [4], etc. Since power analysis attack, in terms of experimental means, only requires passive acquisition of power traces, do not cause any interference with the operation of the cryptographic circuit, and do not require physical damage to the chips as intrusive attacks typically do [5], it is not easily detected. Power analysis attack poses a serious threat to the security of cryptographic devices. Therefore, it is very important to detect whether there is information leakage during the implementation of cryptographic algorithms and to evaluate the ability of cryptographic chips to resist power analysis attack. Power leakage assessment can help designers have a preliminary understanding of the security level of the cryptographic chip and carry out targeted protection, which can greatly improve the security of the cryptographic chip. In addition, a set of sample points on power traces that are most relevant to the sensitive intermediate value of the cryptographic algorithm can be obtained through side-channel leakage detection, and then all sample points on power traces are divided into two groups: leakage points and nonleakage points [6]. Subsequently, the leakage points can be selected as feature points containing the most key-dependent information [7] for power analysis attack, which can effectively reduce the computational complexity and time consumed in the attack and improve the efficiency of the SCA attack.

1.1 Related Work

The main method of cryptographic chip information leakage detection is hypothesis test, based on the principle that intermediate values generated by chips during cryptographic operations can result in power information leakage if intermediate values affect side-channel information(i.e., power traces) in a statistically significant manner. Hypothesis test methods have become the primary means of leakage detection by calculating test statistics and significance levels to identify outliers. For example, Goodwill et al. [8] used a specific t-test to evaluate the side-channel information leakage of the implementation of the AES (Advanced Encryption Standard) encryption algorithm, proving the availability of the t-test in leakage detection. The method divides the collected power traces into two sets based on the individual bit values of the target intermediate values during the operation of the algorithm. If the sample sizes of the two sets are not equal, the Welch's t-test is used to measure the difference between the mean values of the power consumption data of the two sets at each sample point. When the statistic exceeds a threshold at a certain place, it is determined that the traces in two sets have a high confidence difference at that sample point and there is side-channel power information leakage, i.e., the corresponding sample point on the energy curve is the leakage point and the chosen intermediate state value significantly affects the power consumption of the cryptographic chip. In 2013 BECKER et al. [9] proposed the TVLA (Test Vector Leakage Assessment) method, which uses the non-specific t-test to divide the power traces into two groups according to fixed plaintext and random plaintext, and if the value of the t-statistic at a certain time sample exceeds the threshold, that point is a leakage

point. The specific t-test can offer higher confidence in detection at a lower cost than the non-specific t-test [10]. Other common hypothesis test methods include the F-test, which is similar in principle to the t-test in that it detects leakage through statistical differences in power consumption values at every sample point, but the t-test focuses on differences in means between two sets of samples, whereas the F-test focuses on differences in variances.

In addition to the hypothesis test in statistical inference, mutual information is also a commonly used leakage detection method. Mather et al. [11] compared t-test, discrete mutual information (DMI) and continuous mutual information (CMI), in terms of leakage detection capability and computational complexity. The results of the experiment show that the t-test is better when the overall population is normally distributed and the significance of the difference between means needs to be measured. However, CMI can be applied to leakage measurement in any situation, and there are no special requirements for distribution and statistic characteristics.

The t-test is still the most commonly used detection method for side-channel leakage detection. The Welch's t-test is an extension of Student's t-test and is more reliable in cases where the sample sizes of two sets are not equal. Since the Welch's t-test measures the mean difference between two sets of normally distributed populations, in order to accurately use the Welch's t-test to evaluate side-channel leakage information, the key lies in grouping all traces, that is, constructing suitable two sets. At present, there are two main grouping methods for Welch's t-test. The first is to divide all traces into two sets according to fixed plaintext and random plaintext, and the second is to group according to the difference of intermediate variable values. In [12], Welch's t-test is used to measure the side-channel leakage information during the operation of the 3DES algorithm. When constructing the trace sets, only the second type of data is constructed, and eight kinds of variable values are selected from different intermediate states according to the characteristics of the 3DES algorithm. The results of leakage information tests obtained by using Welch's t-test under different groups are different. It can be seen that in the process of using Welch's t-test to detect leakage, the selection of test positions is very important. For the second method of grouping according to different values of intermediate variables, there is a lack of research on specific grouping methods, especially how to group according to bits and Hamming weights.

1.2 Our Contribution

There are two main types of grouping methods for Welch's t-test. The first is to divide into two sets of leakage traces according to fixed plaintext and random plaintext, and the second is to group according to different values of intermediate variables.

In this paper, for the second type of grouping method, the grouping position is determined based on the key intermediate state of the AES algorithm's first round S-box output, combined with the Hamming weight model. Two grouping methods are proposed to measure the leakage information during the operation of the AES algorithm, and the effects of the two methods are compared through experiments. The results show that both the number of leakage points and the distribution ratio of high t-statistic are higher than the method of grouping by bit value.

1.3 Structure of This Paper

The structure of the article is as follows. The first chapter is the introduction, which mainly introduces the research background and research status. The second chapter describes the basic principles of AES and Welch's t-test. The third chapter introduces the selection of test positions and different grouping methods in the process of leakage detection using Welch's t-test, and finally explains the necessity of repeated tests. The experimental results and analysis of the DPA contest V4 dataset are introduced in Sect. 4. In the last chapter, the full text is summarized.

2 Preliminaries

2.1 AES

The Advanced Encryption Standard is the most widely used symmetric encryption algorithm today, which is based on the SP cryptographic structure. Although the increase in the length of the key increases the strength of the algorithm, the number of iterations also increases accordingly. The key length of 128 bits for AES which requires 10 iterations is sufficient for most purposes. In AES encryption, except for the last round, each round of transformation includes 4 basic operations: SubBytes, Shift Rows, Mix Columns, and Add Round Key. There is no Mix Columns operation in the last round.

Shift Rows, Mix Columns and Add Round Key in the AES algorithm are just linear transformations which can only serve as an overall diffusion in the encryption and decryption process. SubBytes is a non-linear transformation, and non-linear transformation is the essence of modern cryptographic algorithms. SubBytes is implemented by replacing each byte in the original matrix with a value obtained from the non-linear component S-box, which has a confusing and local diffusion effect on the data during the encryption process, i.e. a bit different in the S-box input will result in multiple bits different in the output. The good non-linear characteristic can effectively resist traditional cryptanalysis techniques such as linear analysis and differential analysis. Since the S-box transformation is the only nonlinear transformation of the AES algorithm, it largely determines the security of the block cipher. However, its nonlinear characteristic makes it exploited by attackers during side-channel attacks, so the power analysis attack generally selects the output of the first round of S-box or the input of the last round of S-box when the AES algorithm runs as the target intermediate value to carry out the attack.

2.2 Welch's t-test

Welch's t-test is an extension of Student's t-test, which is applicable when the sample size and variance of the two sets are not equal. For n power traces L_i, each containing m sampling points, a certain intermediate value during the operation of the cryptographic algorithm is selected as the grouping basis, and the side-channel leakage traces L is divided into two groups (i.e. L_0 and L_1) according to the two possibilities of this value. The sample sizes, means and variances of L_0 and L_1 are $\left(n_0, \mu_0, S_0^2\right)$ and $\left(n_1, \mu_1, S_1^2\right)$, respectively. The null hypothesis is that the two sets of power consumption traces have

the same mean, and it can be considered that the power traces in the two subsets are not statistically different with high confidence, that is, there is no leakage of side-channel power information. The t-statistic can be expressed as follows.

$$t = \frac{\mu_0 - \mu_1}{\sqrt{\frac{S_0^2}{n_0} + \frac{S_1^2}{n_1}}} \tag{1}$$

Its degree of freedom v is expressed as follows:

$$v = \frac{\left(\frac{S_0^2}{n_0} + \frac{S_1^2}{n_1}\right)^1}{\frac{\left(\frac{S_0^2}{n_0}\right)^2}{n_0 - 1} + \frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1}} \tag{2}$$

The probability density function of the t distribution can be obtained from the degrees of freedom as:

$$f(t, v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}} \tag{3}$$

$\Gamma(\cdot)$ is the gamma function. The probability that the null hypothesis holds in the Welch's t-test is as follows.

$$p = 2\int_{|t|}^{\infty} f(t, v)dt = 2F(-|t|, v) \tag{4}$$

F is the distribution function, which can be expressed as:

$$F(t, v) = \frac{1}{2} + t\Gamma\left(\frac{v+1}{2}\right)\frac{2^{F_1\left(\frac{1}{2}, \frac{v+1}{2}, \frac{3}{2}, -\frac{x^2}{v}\right)}}{\sqrt{\pi v}\Gamma\left(\frac{v}{2}\right)} \tag{5}$$

2^{F_1} is hypergeometric function in (5).

2.3 Pass/Fail Criteria

The probability that the null hypothesis is established in the t-test, that is, the calculation result of Eq. (4), gives the probability that the data mean values of the two sets of power traces at a certain sample point are different. If there is a high-confidence difference between two sets of power consumption values at a particular point, that is, the p-value of the corresponding point is very small, it means that the null hypothesis is rejected and the leakage can be detected. If the p-value is large, it means that the mean difference between the two sets of power consumption data is small, so the null hypothesis is acceptable. The size of the p-value is extremely related to the acceptance or rejection of the null hypothesis, so an appropriate threshold needs to be set. When less severe leaks

need to be detected, the threshold can be set to a small value. Conversely, if a higher level of leakage needs to be detected, the threshold should be set to a larger value. Because of the large computational volume and the complexity of the test process for conducting a full t-test, the actual test is often simplified to calculate only the t-statistic. t-test is usually set at 4.5, as shown in Eq. (6). When the sample size is greater than 1000, setting the threshold for accepting or rejecting the null hypothesis at 4.5 will enable the t-test to be accurate at 99.999% or more [8].

$$p = 2F(t = \pm 4.5, v > 1000) < 10^{-5} \tag{6}$$

When the absolute value of t calculated by Eq. (1) is greater than the threshold value of 4.5, the difference between the two sets of power consumption values is determined, i.e. the leakage of power information can be detected.

3 Leakage Detection with Welch's t-test

3.1 Dataset

This paper conducts experimental analysis and validation with the help of power consumption data from the DPA Contest dataset[13], where the power traces are obtained from an AES-256 RSM (Rotating S-box Masking) implementation. The output of the first S-box which is the first round of the AES encryption algorithm was chosen as the object of the t-test. The 20,000 power traces of the AES encryption algorithm in the DPA Contest V4 dataset have the same fixed key. There are 400,000 samples on each power trace. The choice was to evaluate the first byte of the key for leakage information.

3.2 Welch's t-test Grouping Construction

Most of the existing differential power attacks have been carried out by using the output of the first S-box or the input of the last S-box of the block cipher algorithm run as the target intermediate value. From Eq. (1), it can be seen that the Welch's t-test also uses the difference characteristic. Therefore, in the process of detecting the leakage of power information with Welch's t-test, the first round of S-box output or the last round of S-box input can also be set as the position of the test to improve efficiency. Combined with the analysis of SubBytes in Sect. 2.1, this paper selects the output of the first round of S-box in the AES algorithm encryption process as the sensitive intermediate variable value in the Welch's t-test, proposes two grouping methods, and then constructs two data sets according to the different values of the test position. As shown in Table 1, the first method in the table is the Hamming weight model, and select the Hamming weight (HW) for the first round of S-box output for classification, and the second method is to select each bit of the output value of the S-box for grouping (Fig. 1).

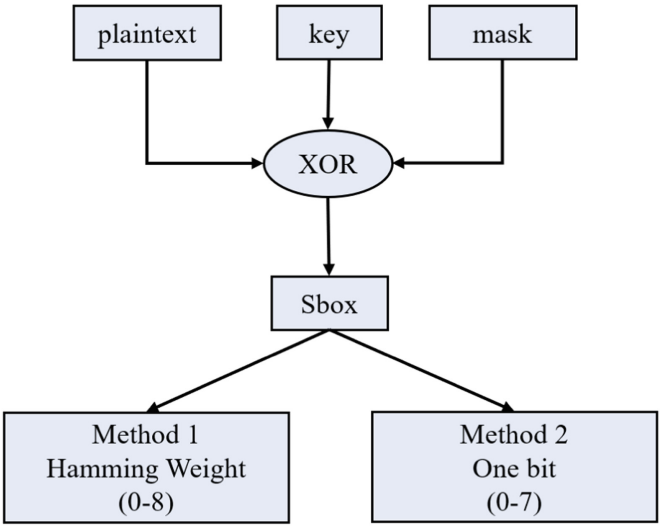

Fig. 1. Diagram of the S-box input and the two corresponding grouping methods

Table 1. Group construction methods for Welch's t-test

	Group method	Set I	Set II
Method I	HW for the first round of the S-box output	$HW = \{0, 1, 2, 3, 4\}$	$HW = \{5, 6, 7, 8\}$
Method II	The first 8 bits values of the S-box output	bit = 0	bit = 1

3.3 Repeated Tests

In the experiment, each leakage trace contains 400,000 sample points. Even if the threshold is set to 4.5, the accuracy of the t-test can reach more than 99.999%, it still cannot be ruled out that at a certain sample point, the absolute value of the t-statistic exceeds the threshold, and there is still a possibility of error at a certain point. In order to minimize the false alarm rate, it is therefore necessary to carry out two independent experiments as a repeat test. 20,000 curves with the same key were selected in DPA contest V4, 10,000 of which were subjected to the Welch's t-test described in Sect. 3.2, and the remaining 10,000 were subjected to the same test. For a given sample point on the energy curve, the sample point can only be considered a leakage point if the threshold is exceeded at the same time in both experiments, because if it is chance that the t-statistic exceeds the threshold at a given time, it is unlikely that this will be repeated at the same time in the next repeat experiment.

Combining the content of Sect. 3.2, a flow for side channel leakage assessment using the Welch's t-test can be obtained as shown in the figure below (Fig. 2).

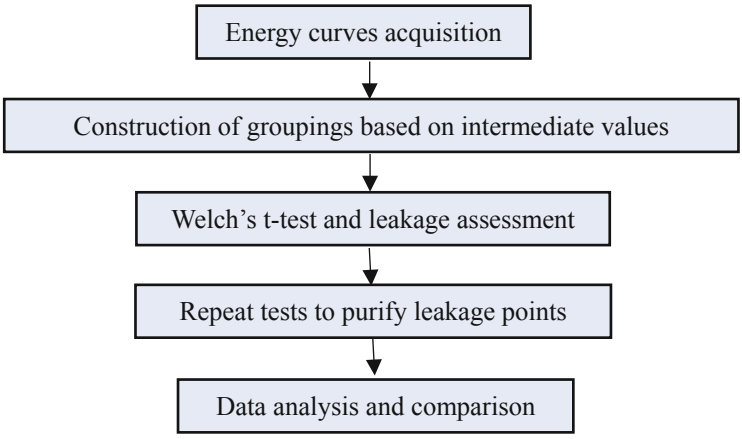

Fig. 2. Schematic of the flow of side-channel leakage assessment with Welch's t-test

4 Experimental Results and Analysis

In DPA contest V4, select 20,000 energy curves with the same key, of which 10,000 are subjected to Welch's t-test according to the two grouping methods described in Sect. 3.2, and the remaining 10,000 energy curves are repeated according to Sect. 3.3. The sample points which exceed the threshold in both experiments are leakage points, and the number of leakage points detected by each grouping method is shown in Fig. 3. It can be seen from the figure that each grouping method can detect a large number of leakage points. But in comparison, the method that selects the Hamming weight output by the first round of S-boxes for grouping can detect 1002 leakage points. It can detect the larger number of leakage points. In the second method, that is, the method of grouping according to the bit value of 0 or 1, the third bit and the seventh bit outperform the other 6 grouping methods, and can detect 956 and 932 leakage points respectively.

Figure 4 shows the result of Welch's t-test for the first group of 10,000 curves by method I in Table 1, i.e., the Hamming weight model. The red dotted lines in the two figures represent the threshold ± 4.5, and it can be found that the leakage points are mainly distributed in four regions, i.e., the four peaks in the figures. In Fig. 5, the leakage points obtained by grouping with the method II are also distributed in the same four main areas, and the $|t|_{\max}$ obtained by each grouping method in each area and the corresponding sample time points are recorded in Table 2. It can be seen from the data in the table that although the corresponding leakage time points of $|t|_{\max}$ in the four main areas are similar for each method, there is still a certain gap in the t-statistic. The t-statistic of HW is generally larger in all four regions. The t-statistic of bit7 is larger in the other three regions but smaller in region IV. Figure 6 zooms in on the three areas of the area marked by the red square in Fig. 4 to obtain a local curve graph of the area. The figure again confirms that the t-statistic obtained by the HW method at most sample points are higher than those obtained by the other seven grouping methods.

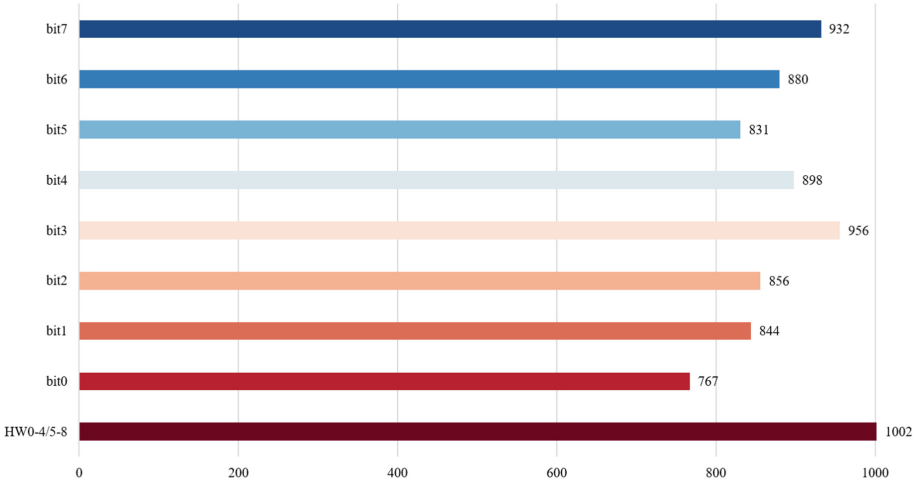

Fig. 3. The number of leakage points obtained by each grouping method

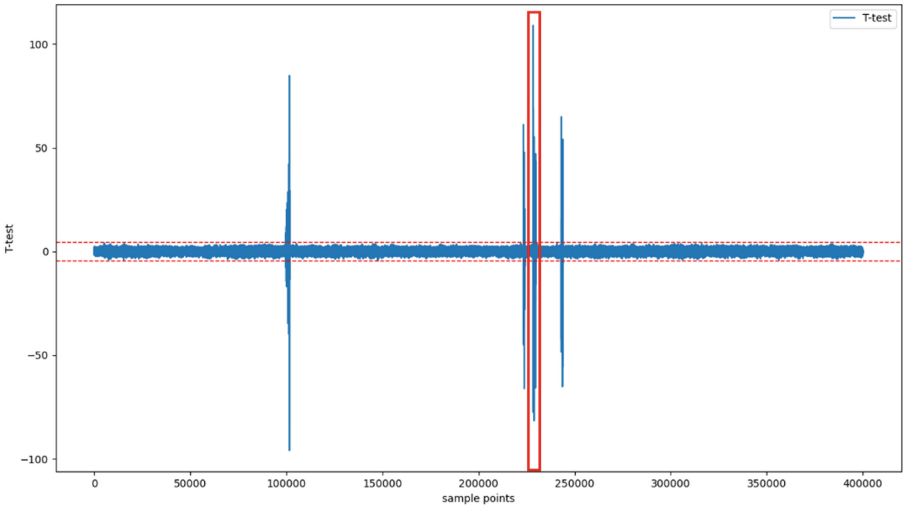

Fig. 4. T-statistic at different sample points obtained by method I

Table 2 only lists the distribution of the maximum t-statistic in the four regions. In order to analyze the distribution of the magnitude of t-statistic at all leakage points under different grouping methods, Fig. 7 divides the magnitude of t- statistic into six groups, $4.5 \leq t < 10$, $10 \leq t < 20$, $20 \leq t < 30$, $30 \leq t < 40$, $40 \leq t < 50$ and $t \geq 50$, respectively. Although the maximum t-statistic of 156.625 obtained by grouping by bit7 values was the highest of all results, the t-statistic of leakage points obtained by grouping by first round S-box output Hamming weights of 0–4 or 5–8 were the most evenly distributed in the six intervals, especially in the intervals $30 \leq t < 40$, $40 \leq t < 50$ and $t \geq 50$,

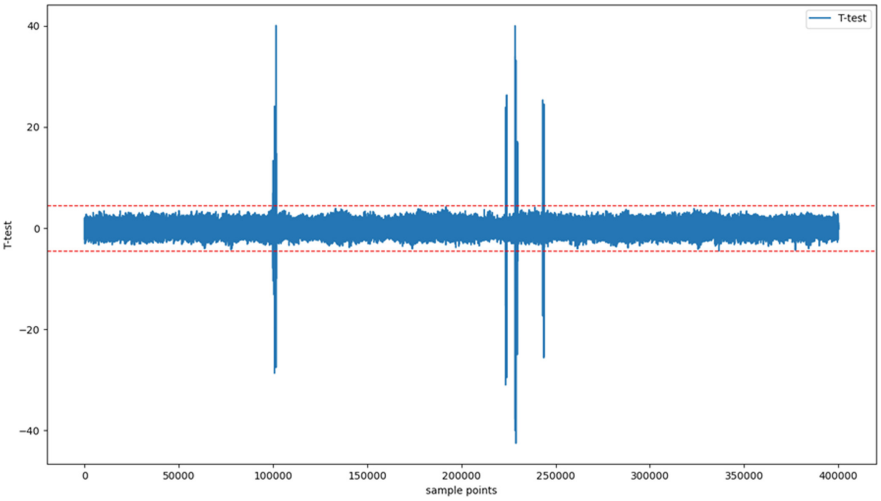

Fig. 5. T-statistic at different sample points obtained by bit0

Table 2. $|t|_{\max}$ and its corresponding leakage point in four regions under different grouping methods

	Region I $\|t\|_{\max}$ (sample point)	Region II $\|t\|_{\max}$ (sample point)	Region III $\|t\|_{\max}$ (sample point)	Region IV $\|t\|_{\max}$ (sample point)
bit0	39.921(101577)	30.179(223980)	42.353(228894)	25.371(243104)
bit1	38.414(101575)	27.262(223779)	55.533(228414)	28.066(243666)
bit2	40.916(101577)	28.452(223977)	55.001(228416)	24.980(243110)
bit3	49.152(101580)	40.444(224106)	49.623(228395)	34.141(243112)
bit4	57.329(101435)	41.766(224108)	48.788(228590)	30.694(243108)
bit5	31.926(101590)	29.579(224107)	36.410(228470)	25.136(243668)
bit6	70.178(101578)	54.011(224110)	49.812(228392)	37.492(243665)
bit7	156.625(101571)	101.138(224112)	76.217(228593)	25.223(243668)
HW	94.485(101589)	64.570(223781)	108.256(228403)	63.697(243108)

where the number of leakage points was most prominent and much higher than those obtained by other methods. The size of the t-statistic represents, to a certain extent, the degree of leakage, i.e., the degree of correlation with the sensitive intermediate value. When doing subsequent first-order DPA or CPA attacks, a portion of the leakage points can be selected as feature points according to the number of feature points required, using the size of the t-statistic as a reference standard. In summary, grouping by AES first round S-box output Hamming weight of 0–4 or 5–8 works better than grouping by S-box output of 0 or 1 for each bit value.

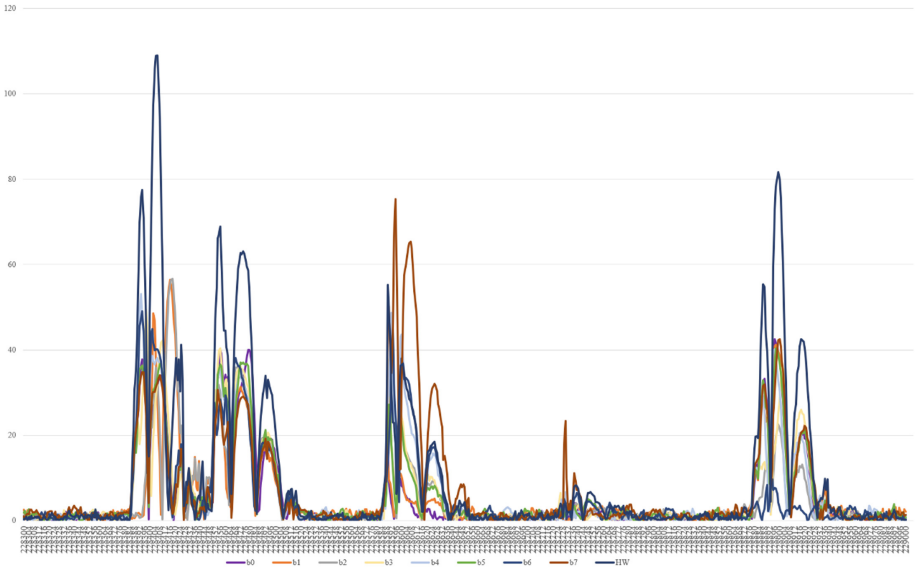

Fig. 6. T-statistic at several sample points obtained by every grouping method for region III

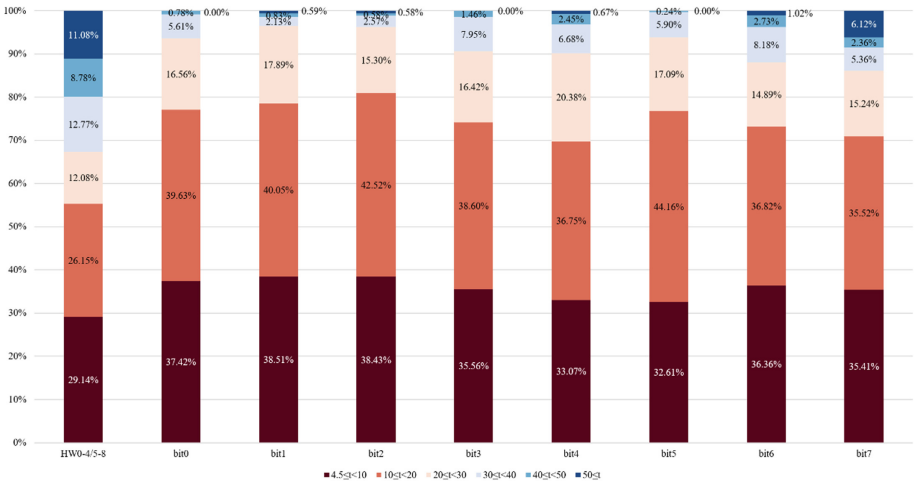

Fig. 7. Distribution of the magnitude of t-statistic for all leakage points

5 Summary

This paper examines the choice of test positions in the process of side-channel information leakage assessment using the Welch's t-test, and proposes two grouping methods. Experiments on the DPA contestV4 dataset lead to the conclusion that the method of grouping by first-round S-box output Hamming weight of 0–4 or 5–8 is higher than the

other grouping methods in terms of both the number of leakage points and the proportion of high t-statistic distribution.

The next step can be to use the side-channel leakage assessment method proposed in this paper to compare it with other tests in the field of statistics such as the F-test and the chi-square test. In addition, a study can be conducted to verify whether the higher the t-statistic of the leakage point is chosen as the feature point, the higher the success rate of the first-order side-channel attack will also be.

Acknowledgments. This research was supported by the College Students' Innovation and Entrepreneurship of china (No. 202210018009).

References

1. Mangard, S., Oswald, E., Popp, T.: Power Analysis Attacks: Revealing the Secrets of Smart Cards. Springer, Berlin (2007). https://doi.org/10.1007/978-0-387-38162-6
2. Kocher, P., Jaffe, J., Jun, B.: Differential power analysis. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48405-1_25
3. Brier, E., Clavier, C., Olivier, F.: Correlation power analysis with a leakage model. In: Joye, M., Quisquater, J.-J. (eds.) CHES 2004. LNCS, vol. 3156, pp. 16–29. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28632-5_2
4. Chari, S., Rao, J.R., Rohatgi, P.: Template attacks. In: Kaliski, B.S., Koç, çK., Paar, C. (eds.) CHES 2002. LNCS, vol. 2523, pp. 13–28. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-36400-5_3
5. Bao, S.G.: Research on the Experimental Method and Technology of Smart Card Template Attack. Shanghai Jiaotong University, Shanghai (2015)
6. Bhasin, S., Danger, J., Guilley, S., et al.: NICV: normalized interclass variance for detection of side-channel leakage. In: Proceedings of the EMC 2014, Tokyo, Japan, pp. 310–313 (2014)
7. Chen, S., Rui, W., Wang, X.F., et al.: Side-channel leaks in web applications: a reality today, a challenge tomorrow (2010)
8. Goodwill, G., Jun, B., Jaffe, J., et al.: A testing methodology for side-channel resistance validation. In: Proceedings of the NIST NIAT 2011, Nara, Japan, pp. 115–136 (2011)
9. Becker, G., Cooper, J., Demulder, E., et al.: Test vector leakage assessment (TVLA) Methodology in Practice. http://pdfs.semanticscholar.org/a10f/31018c9ce38a5231b6481a8f9d4881bca64c.pdf, Accessed 30 Apr 2020
10. Schneider, T., Moradi, A.: Leakage assessment methodology. In: Güneysu, T., Handschuh, H. (eds.) CHES 2015. LNCS, vol. 9293, pp. 495–513. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-48324-4_25
11. Mather, L., Oswald, E., Bandenburg, J., Wójcik, M.: Does my device leak information? an a priori statistical power analysis of leakage detection tests. In: Sako, Kazue, Sarkar, Palash (eds.) Advances in Cryptology - ASIACRYPT 2013, pp. 486–505. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42033-7_25
12. Chen, J., Li, H., Wang, Y., Wang, Y.: Evaluation side-channel information leakage in 3DES using the t-test. J. Tsinghua Univ. (Nat. Sci. Ed.) **56**(05), 499–503 (2016). https://doi.org/10.16511/j.cnki.qhdxxb.2016.25.007
13. TELECOM ParisTech SEN research group. DPA Contest (4th edition) 2013–2014. http://www.DPAcontest.org/v4/