MIDI: A Fast Method of Protocol Format Extraction based on Mutual Information Difference

Zhengguo Xu*†, Hui Zheng†, Jiaqi Yao†, Keren Wang†
* University of Electronic Science and Technology of China
Email: zhg.xu.ac@gmail.com
† National Key Laboratory of Science and Technology on Blind Signals Processing, China

***Abstract*—Protocol reverse analysis is an important issue in network security research, in which the automatic extraction of protocol format is the critical step. The existing methods mainly leverage the multiple sequence alignment or the NLP models to extract the protocol format. A major drawback of these methods is that the high computational complexities limit their practicalities. To tackle this problem, we propose a protocol format reverse method based on mutual information difference, called MIDI. The computational complexity of MIDI is linear with the size of protocol samples, in terms of that the accuracy of field segmentation is improved proportionally when the sample size increases. The basic idea of MIDI comes from the difference of statistical characteristics between neighboring fields, and MIDI requires only the local alignment of the fields, not the entire protocol format alignment. Based on the statistical properties of protocol fields and the key factors of indicating the field boundary, we give a theoretical analysis of the field delimitation conditions. Further, we define a new measurement for evaluating the accuracy of format extraction. We select nine protocols in diverse network layers, including seven known protocols and two private protocols whose formats have been disclosed manually, as ground truth for testing. The reversing performances achieve over 30% without any prior knowledge of the protocol specifications, especially, the accuracy of reversing TCP format is 100%.**

I. Introduction

Protocol format reverse analysis aims to extract the fields of undocumented protocols and identify the field structure. It is applied in many security contexts, such as malicious traffic identification, protocol vulnerability mining and network intrusion detection. The traditional protocol analysis relies on manual extracted features. To avoid the error-prone and involved process, a number of trace-based approaches using machine learning algorithms were proposed in recent years, such as multiple sequence alignment, frequent subsequence mining, n-gram, voting experts, and hidden semi-Markov model [1], [2], [3], [4], [5], [6].

Although these studies claim the effectiveness of their methods, their computational complexities are $O(n^2)$ or even higher, where n is the number of protocol samples. Besides, some analysis processes need prior knowledge of protocols, such as the known delimiters or tokens. The innovation of this paper is that we discuss the delimitation conditions of protocol fields in different cases and introduce the mutual information to complete the field segmentation. We propose a format extraction method called MIDI, which decreases the computation complexity to $O(n)$ without the process of mining sequence patterns, and does not require the entire protocol format alignment or any prior knowledge. Further, we define a new measure of reverse accuracy to evaluate MIDI's performance. It offers a two-fold description of the field: the boundary deviation and the range fuzziness. In the experiments, we take the known protocol formats as ground truth to validate the proposed method. There are nine protocols in different network layers, including seven known protocols and two private protocols whose formats have been disclosed manually. The results show that the format extraction accuracy is over 30%, especially, the performance of reversing TCP format can reach 100%.

II. Methodology

The intuition behind our proposed methods is to explore the correlation between the field boundary and the differences in statistical properties of neighboring fields. More precisely, let the discrete random variables X, Y denote the values of two neighboring fields, and suppose that X, Y follow independent probability distributions respectively. When X, Y are serialized as a byte string, $X := X_0X_1 \dots X_{n-1}, Y := Y_0Y_1 \dots Y_{m-1}$, there is a bijection $f : \mathbb{R}^n \to \mathbb{R}$ that makes $X = f(X_0, X_1, \dots, X_{n-1})$, since the field serialization is a reversible and unique process ruled by its protocol specification. According to the independence assumption of X, Y and the bijection f, we discuss the field delimitation conditions in detail below. Let

$$Z := Z_0Z_1 \dots Z_{n-1}Z_nZ_{n+1} \dots Z_{n+m-1}, \tag{1}$$

$$Z_k = \begin{cases} X_k, & k \in [0, n); \\ Y_{k-n}, & k \in [n, m). \end{cases} \tag{2}$$

And we define $Z_{\leq k} := Z_0Z_1 \dots Z_k$. The goal is to find the position k in the byte sequence, making $Z_{\leq k-1} = X$. Denote

$$\alpha_k = \begin{cases} H(Z_0), & k = 0; \\ \beta_{k-1} + H(Z_k), & k \in [1, n+m), \end{cases} \tag{3}$$

$$\beta_k = H(Z_{\leq k}), \qquad k \in [0, n+m), \tag{4}$$

where H is the information entropy. From the properties of H we have,

$$I_k := I(Z_{\leq k-1}; Z_{\leq k}) = \alpha_k - \beta_k. \tag{5}$$

978-1-5090-6501-1/17/$31.00 © 2017 IEEE

In equation 5, I_k is the average mutual information between $Z_{\leq k-1}$ and $Z_{\leq k}$, scilicet, it is the average amount of information that the k-th byte obtained from the pre-sequence $Z_{\leq k-1}$.

Then we have the following theorem.

Theorem 1 *$I_k - I_{k-1} \leq 0$ is a necessary but non-sufficient condition for determining the field boundary at the k-th byte.*

In practice, the protocol field can be divided into three types, the constants, the random variables subjected to uniform distribution and the random variables subjected to non-uniform distribution. Due to the constraint of protocol specification, the value of protocol field exhibits a stable distribution in a sufficient sampling. We apply the above theorem to analyze the neighboring cases of different field types, then we propose MIDI, a new format extraction method.

III. The Algorithm of MIDI

Applying MIDI to extract the protocol format, it requires the fields to be aligned in the samples. For the fixed format protocols, the fields are inherently aligned. For the dynamic format protocols, the positions of some fields may be floated in different samples. We use the locality-sensitive hashing to get the similar subsequences and apply multi sequence alignment to match the bytes. For the cases of two neighboring variables subjected to non-uniform distributions, the boundaries can be revealed by computing the turning points of I_k recursively. For the constant cases, we can check the change points of I_k.

Figure 1 shows the result of reversing the TCP header format. The green dotted lines represents the real field boundaries, the red solid lines is the boundaries deduced by MIDI. The results show that MIDI identifies all fields in TCP header. Since MIDI depends on the probability distribution of the field values, Figure 2 illustrates the accuracy of reversing TCP header format improved by the increasing samples.

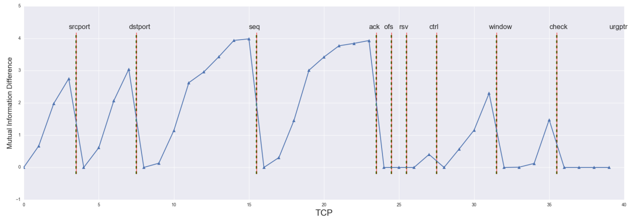

Fig. 1. The segmentation results of TCP format

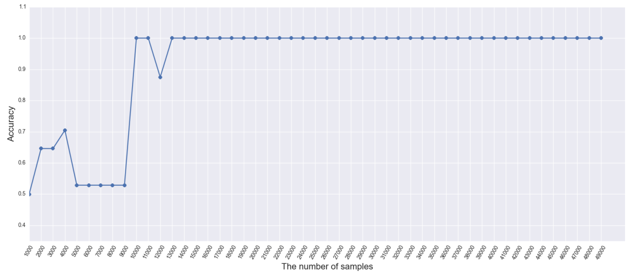

Fig. 2. The performance of MIDI improved by the increasing samples

TABLE I. The format extraction results of MIDI

Protocol	Layer	Acc	Fuz	Dev
ip	network	0.47	0.25	0.75
tcp	transport	1	0	0
ntp	application	0.37	0.04	0.63
nbns	application	0.3	0.35	0.54
stun	tunneling	0.32	0	0.58
a11 reply	session	0.36	0.3	0.69
a11 ack	session	0.47	0.27	0.69
private 1	application	0.37	0.30	0.61
private 2	application	0.35	0.30	0.64

IV. Evaluation

To evaluate the performance of reversing format, we define a new accuracy measure. It consists of two parts, the boundary deviation and the range fuzziness. It is defined as follow.

$$Acc = \left(1 - \frac{Dev + Fuz}{2}\right) \cdot \log\left(\frac{|m-n|}{n} + \mathrm{e}\right)^{-1}, \qquad (6)$$

where n is the true number of protocol fields, m is the number of fields inferred by MIDI. In equation 6, Dev evaluates the displacement between the real boundary and the reversed boundary, while Fuz measures the coverage of the reverse field on the real field ranges. The $\log(\cdot)$ term is the punishment, the greater the difference between m and n, the lower the accuracy. Table I shows the format extraction results of MIDI. In addition to TCP, we examine our method by more protocols, including two private protocols whose formats have been disclosed, and other protocols in different layers. The accuracies of reversing these protocols are more than 30%.

V. Discussion

Comparing with the existing research work, our approach does not require protocol data to be strictly aligned in the field format, and we only scan protocol data once to complete the field segmentation. The processing efficiency is 1000 times that of other methods when we deploy MIDI on Spark. But there are two on-going improvements. The false positive boundaries could be eliminated when we improve the segmentation criteria, and we are verifying the performance and applicability of MIDI on other more complex protocols, such as DNS.

References

[1] J. Narayan, S. K. Shukla, and T. C. Clancy, “A survey of automatic protocol reverse engineering tools,” *ACM Comput. Surv.*, vol. 48, p. 40, 2015.

[2] X. D. Li and L. Chen, “A survey on methods of automatic protocol reverse engineering,” in *CIS*, 2011.

[3] Z. Zhang, Z. B. Zhang, P. P. C. Lee, Y. J. Liu, and G. G. Xie, “ProWord: An unsupervised approach to protocol feature word extraction,” in *Proc. of IEEE INFOCOMM*, 2014.

[4] J. Antunes, N. F. Neves, and P. Veríssimo, “Reverse engineering of protocols from network traces,” in *WCRE*, 2011.

[5] J. Z. Luo and S. Z. Yu, “Position-based automatic reverse engineering of network protocols,” *J. Network and Computer Applications*, vol. 36, pp. 1070–1077, 2013.

[6] M. Li and S. Z. Yu, “Noise-tolerent and optimal segmentation of message formats for unknown application-layer protocols (in chinese),” vol. 24, no. 3, pp. 604–617, 2013.