A Format Reverse Method for Binary Protocol from Communication Data

Fanzhi Meng, Yuan Liu, Chunrui Zhang, Dong Liu

Institute of Computer Application, China Academy of Engineering Physics, Mianyang, China

mengfz@caep.cn

Keywords: protocol format reverse; HMM; multiple sequence alignment; feature analysis

Abstract. Protocol format reverse based on communication data has played an important role in the fields of network security and information countermeasures. In this paper, a format reverse analysis method for binary communication protocol which based on probability alignment and differential analysis of statistic is proposed. The method adopts the data set of protocol frame as analysis object, and makes the corresponding fields in protocol frame aligned accurately by probability alignment algorithm firstly, and then identifies the boundary of adjacent fields in the frame according to the different features of various statistics, and finally reverses the communication protocol format specification. The experimental results show that the method can effectively identify the format specification of binary communication protocol and semantics specification for some fields in protocol frame format.

Introduction

With the development of the Internet, the security situation of network becomes more severe. Protocols are a foundational aspect of network and hence it is important for security analysts to understand the protocol. However, it is very difficulty for security analysts understand the undocumented protocol. Communication protocol reverse engineering appeared for obtaining the document of undocumented protocol, and mining the protocol specification from the communication data become one of the focuses of communication protocol reverse engineering [1].

A complete communication protocol frame usually consists of the protocol data from link layer to application layer and the information data. The underlying protocols are usually the binary protocol, and the application layer protocols are mostly text protocol. So communication protocol frame format can be classified into binary format and text format [2]. The present studies on protocol format reverse analysis mainly focus on text protocols, and lack of studies on binary protocols.

The format reverse engineering of unknown protocol based on communication data has been widely used in the fields of network security, such as deep packet detection [3], fuzzy test [4] and many other technologies. However, most of the studies [5-11] focus on text-based protocols or application layer protocols based on Ether network environment, which will not apply to binary protocols, such as the Discoverer [5], ReverX [6], ProDecoder [7], etc. PI [8] which uses gene sequence alignment for reference, can only determine some fields from the data of the protocols with simple formats and conservative sequences.

In this paper, the format reverse analysis method for binary communication protocol from communication dada which combined with probability alignment and differential analysis of statistic is proposed. The method takes the protocol frame data as the analysis object, and makes the corresponding fields in protocol frame aligned accurately by probability alignment algorithm firstly, and then various statistics of fields are calculated, and finally identifies the boundary of adjacent fields in the frame and the format of protocol according to the different features of various statistics. To validate the method, the Internet communications data is used for experimental data. The experimental results show that the method can effectively identify the format specification of unknown binary communication protocol.

Architecture of the System

Theoretical basis.

Protocol specification is an agreement which must be complied with by the protocol entities in the process of packet construction and transmission, so the construction and transmission of packet are restricted by the protocol specification. The constraints or characteristics of protocol specification will highlight when communication data is large enough. This is why we can reverse part of protocol specification from communication data. We summary the following two general principles by the analysis of a lot of protocols:

1) The format of a packet is not randomly chosen and that a structure is defined in order to allow a communication between two entities.

2) The packet of one protocol with a specific version does not change even when the communication is repeated at different location and times.

System framework.

The framework of the unknown protocol format reverse analysis system is shown in Fig.1. According to the implementation process, it can be divided into three parts include fields alignment, statistics calculation and format identification.

The essential of the protocol format reverse is the separation of different fields in the protocol frame. The theoretical basis used in this paper is that the characteristics of the values of adjacent fields (e.g., statistical features, distribution features) would be different, and so can be used for the division of adjacent fields. To highlight the characteristics of fields for statistical analysis, the corresponding fields in the frames must be aligned. Progressive multiple sequence alignment algorithm has good effects in the field of biological gene sequence alignment. In order to adapt to the characteristics of binary protocol frame data, the paper design a probability alignment algorithm based on HMM for field alignment process and it can improve the accuracy of alignment of protocol frame fields. After the calculation of various statistics, such as change rate, mean, variance and ratio of major, etc, separating the adjacent fields based on the different value of statistics. At last, we reconstruct the unknown protocol format according to the segmentation of fields and infer the function of some fields in coarse-grained level.

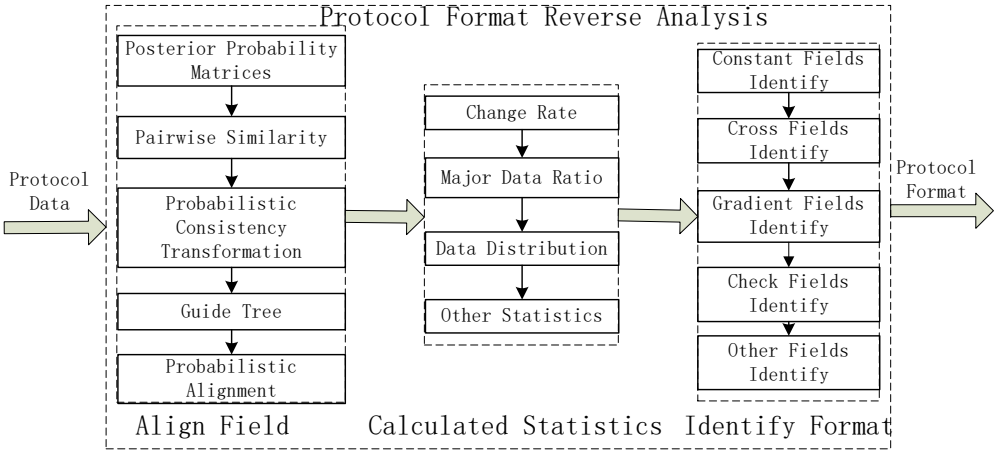

Figure 1 The format reverse system framework of unknown protocol based on probability alignment

Multiple Sequence Alignment Algorithm Based on HMM

Guide tree is an important premise of progressive multiple sequence alignment, but the similarity measurement of two sequences is isolated, which leads to the construction of guide tree is not accurate, and further causes the result of the multiple sequence alignment is not accurate. To reduce the negative impact of deviation caused by guide tree, the reference sequence was introduced into the progressive multiple sequence alignment.

Therefore, the paper introduces the probabilistic consistency method which revises scoring system by providing the consistency probability information by the third reference sequences for two sequence alignment which is based on Pair - HMM model. Analogy to the traditional progressive multiple sequence alignment algorithm, multiple sequence alignment algorithm based on HMM is divided into the following five stages:

1) Posterior probability matrix computation;
2) Pairwise similarity computation;
3) Probabilistic consistency transformation;
4) Guide tree construction;
5) Progressive multiple sequence alignment.

Posterior probability matrix computation.

Let x and y be two frames represented as character strings in which x_i is the ith character of x. Consider the pair-HMM given in Fig.1, where $A_{x,y}$ is the space of all possible x~y alignments.

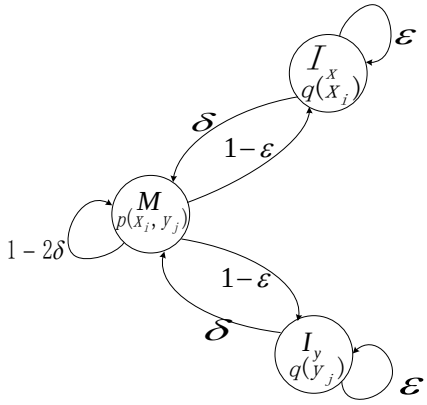

Figure 2 the structure of Pair - HMM

An alignment a corresponds uniquely to a sequence of state-emission pairs, $<s_1, o_1>,...,<s_n, o_n>$. The probability of a is given by:

$$p(a)=\pi(s_1)\left(\prod_{i=1}^{n-1}\alpha(s_i\to s_{i+1})\right)\left(\prod_{i=1}^{n}\beta(o_i|s_i)\right)\pi(s_n) \tag{1}$$

Where $\pi(s)$ is the initial or final probability of state s, $\alpha(s_i \to s_i{+}1)$ is the transition probability, and $\beta(o_i \mid s_i)$ is the emission probability. Let a^* be the alignment from $A_{x,y}$ which most nearly represents the "true" alignment of x and y. We assume that $\mathbf{P}(a \mid x, y)$ is the probability that an alignment a is equeal a^*. Let the notation $x_i \sim y_j \in a$ denote the event that positions x_i and y_j are matched in an alignment a. Formally, the posterior probability of $x_i \sim y_j \in a^*$ is:

$$P(x_i\sim y_i\in a^*|x,y)=\sum_{a\in A_{x,y}}P(a|x,y)1\{x_i\sim y_i\in a\} \tag{2}$$

Where $1\{x_i \sim y_j \in a\}$is the condition function which evaluates to 1 whenever is true and 0 otherwise. Then the posteriori probability matrix P_{xy} for the alignment of x and y is a table of $P(x_i \sim y_j \in a^* \mid x, y)$ values for $1 \le i \le |x|$, $1 \le j \le |y|$.

Pairwise similarity computation.

In this work, we find the alignment a which does not maximize the probability of $a = a^*$ but rather tries to guarantee high accuracy for a, which we define with respect to the alignment a^* as:

$$Acc(a)=\frac{1}{\min\{|x|,|y|\}}\sum_{x_i\,\square\,y_i\in a}1\{x_i\,\square\,y_i\in a^*\} \tag{3}$$

During the alignment process, however, a^* is not known, so we instead maximize the expected accuracy of the reported alignment. And the expected accuracy is adopted to measure the similarity between two sequences. The expected accuracy of the reported alignment is given by:

$$E(Acc(a))=\frac{1}{\min\{|x|,|y|\}}\sum_{x_i\,\square\,y_i\in a}P\{x_i\,\square\,y_i\in a^*|x,y\} \tag{4}$$

Using this decomposition, we compute the maximal expected accuracy alignment by a simple variant of Needleman-Wunsch algorithm [4], where all match/mismatch scores are given by the posterior probability terms for corresponding letters, and gap penalties are set to zero.

Probabilistic consistency transformation.

In the previous subsection, we described a method for performing pairwise sequence alignment of two sequences *x* and *y* based on computing $P(x_i \sim y_j \in a^* \mid x, y)$ values for all positions in *x* and *y*, and subsequently using these posterior probabilities as match/mismatch scores in a Needleman-Wunsch like alignment procedure. In this subsection we introduce probabilistic consistency, a method for obtaining more accurate substitution scores when a third homologous sequence *z* is available.

For a sequence *z*, let $z(k,k+1)$ denote the inter-letter regions (or gaps) between character *k* and $k+1$ of *z* for $0 \le k \le |z|$ (where $z(0,1)$ and $z(|z|,|z|+1)$ denote the gaps at the beginning and ends of *z*). Generalizing our notation for posterior probability of matches, an alternative estimate for the quality of a $x_i \sim y_j$ match is given by marginalized probability:

$$P\left(x_i \sim y_i \in a^* \middle| x, y, z\right) = \sum_{z_k} P\left(x_i \sim y_i \sim z_k \in a^* \middle| x, y, z\right) + \sum_{z_{(k,k+1)}} P\left(x_i \sim y_i \sim z_{(k,k+1)} \in a^* \middle| x, y, z\right) \quad (5)$$

To simplify the calculation, ignored the second summation over gaps in *z*, and applied heuristic independence assumptions to get the follow expression.

$$P\left(x_i \sim y_i \in a^* \middle| x, y, z\right) = \sum_{z_k} P\left(x_i \sim z_k \in a^* \middle| x, z\right) P\left(z_k \sim y_i \in a^* \middle| z, y\right) \quad (6)$$

With the procedure described above, we can align two sequences given information from a third sequence. To align two sequences *x* and *y* given a set of multiple sequences *S*, we would ideally estimate $P(x_i \sim y_j \in a^* \mid S)$. In practice, we use the heuristic decomposition:

$$P\left(x_i \sim y_i \in a^* \middle| S\right) = \frac{1}{|S|} \sum_{z \in S} \sum_{z_k} P\left(x_i \sim z_k \in a^* \middle| x, z\right) P\left(z_k \sim y_i \in a^* \middle| z, y\right) \quad (7)$$

Due to the posteriori probability alignment matrices tend to be spares with most entries near zero, so we can reduce the calculation complexity by ignoring these smaller entries of matrix to zero.

Guide tree construction.

The paper will use a greedy heuristic method reminiscent of UPGMA [5] to construct a tree with high expected alignment reliability. Given a set *S* of sequences to be aligned, denote the expected accuracy for aligning any two sequences *x* and *y* as $E(x, y)$. Initially, each sequence is placed in its own cluster. Then, the two clusters *x* and *y* with the highest expected accuracy are merged to form a new cluster *xy*; we then define the expected accuracy of aligning *xy* with any other cluster *z* as $E(x, y)(E(x, z) + E(y, z))/2$. This process is repeated until only a single cluster remains.

Progressive multiple sequence alignment.

The final progressive alignment step is a routine extension of maximal expected accuracy alignment to an unweighted sum-of-pairs model. For each progressive alignment step, we run a profile-profile Needleman-Wunsch alignment procedure in which the score for matching a column containing n_1 non-gap characters to one with n_2 non-gap characters is computed by summing $n_1 n_2$ values from the corresponding pairwise posterior matrices. Note that no gap penalties are used in this final step, thus greatly simplifying the task of profile-profile alignment.

Protocol Format Analysis and Inference

In the communication process, the protocol data will be generated in strict accordance with the protocol specification constraints. Due to the different fields in protocol format have different constraints caused by the protocol specification, and the constraints will display though the characteristics of statistics of fields when the communication data is enough. This is also the theoretical basis that separating the adjacent fields in protocol frames by the characteristic analysis of these fields. The fields in the frames will be aligned based on the constant fields after the protocol frame collection is processed by multiple sequence alignment based on HMM, and the collection of protocol frames will turns into a two-dimensional frame matrix in which each line is a frame with

some gaps in it. The constraint of values in the same filed of frames is the same in the process of communication, and the constraints between adjacent fields are different. We can make the statistical features of characters in the same domain field are similar and in different domain fields are different obviously by designing appropriate statistics. If the values of the statistics between two adjacent characters are different obviously, the two adjacent characters will be the partition boundary of two domain fields. After the calculation of various statistics of each column in the two-dimensional frame matrix is completed, such as change rate, mean, variance and ratio of major etc, we can merge the adjacent characters with the similar value of statistic into one domain field, and separate the adjacent characters with the different values of statistic into different domain fields. The protocol format segmentation completed when all characters have been divided into one domain field.

On the basis of protocol format segmentation, we can infer the semantics and function of each domain field in the protocol format according to the statistical characteristic. For example, if the change rate of a domain field is close to 100%, thus the domain field may be checksum or serial number field. We can further determine the function type of the domains field through the analysis of distribution of field value.

Experiment and Results

To validate the method of protocol format reverse analysis presented in the paper. We use the Wireshark which can capture the network data to obtain the Internet data as the experimental data, and test the validity of the reverse analysis method without any prior knowledge.

The experimental results and analysis of multiple sequence alignment based on HMM.

The value of certain fields in the frame changes frequently, which causes different fields in two frames have the same value. In order to cater to the special situation, traditional progressive multiple sequence alignment algorithms will insert some gaps into the frames, as shown in Fig.3, so as to make the near field cannot be aligned. Fig.4 is the results obtained by our probability multiple sequence alignment.

From Fig.4, we can see that the multiple sequence alignment based on HMM can eliminate the negative impact of special frames on the entirety result, and improve the accuracy of multiple sequence alignment results, so that the corresponding fields in the protocol frame are aligned correctly.

```
01: 0xAB 0xC6 0x6B 0x0A 0xAC 0x27 ___ ___ ___ 0x36 0xFF 0xC2 0xDC 0x11
02: 0xAB 0xC6 0x6B 0x0B 0xAC 0x27 ___ ___ ___ 0x40 0x35 0xF5 0xDC 0x78
03: 0xAB 0xC6 0x6B 0x0C 0xAC 0x27 ___ ___ ___ 0x4A 0xEA 0xF1 0xDC 0x3A
04: 0xAB 0xC6 0x6B 0x0D 0xAC 0x27 ___ ___ ___ 0x54 0x00 0x01 0xDC 0x02
05: 0xAB 0xC6 0x6B 0x0F 0xAC 0x27 ___ 0x6E 0x67 0x19 ___ ___ 0xDC 0x08
06: 0xAB 0xC6 0x6B 0x11 0xAC 0x27 0x23 0x6E 0x67 ___ ___ ___ 0xDC 0x26
07: 0xAB 0xC6 0x6B 0x12 0xAC 0x27 ___ ___ ___ 0x84 0x04 0x02 0xDC 0x01
08: 0xAB 0xC6 0x6B 0x13 0xAC 0x27 ___ ___ ___ 0x8E 0x41 0x51 0xDC 0x22
09: 0xAB 0xC6 0x6B 0x14 0xAC 0x27 ___ ___ ___ 0x98 0x55 0xB1 0xDC 0x2D
10: 0xAB 0xC6 0x6B 0x15 0xAC 0x27 ___ ___ ___ 0xA2 0xC1 0x88 0xDC 0xC3
11: 0xAB 0xC6 0x6B 0x16 0xAC 0x27 ___ ___ ___ 0xAC 0x2B 0x57 0xDC 0x99
12: 0xAB 0xC6 0x6B 0x17 0xAC 0x27 ___ ___ ___ 0xB5 0x6B 0xFC 0xDC 0x04
13: 0xAB 0xC6 0x6B 0x18 0xAC 0x27 ___ ___ ___ 0xBF 0x17 0x07 0xDC 0xBE
```

Figure 3 The sequences set processed by traditional multiple sequence alignment

```
01: 0xAB 0xC6 0x6B 0x0A 0xAC 0x27 0x36 0xFF 0xC2 0xDC 0x11
02: 0xAB 0xC6 0x6B 0x0B 0xAC 0x27 0x40 0x35 0xF5 0xDC 0x78
03: 0xAB 0xC6 0x6B 0x0C 0xAC 0x27 0x4A 0xEA 0xF1 0xDC 0x3A
04: 0xAB 0xC6 0x6B 0x0D 0xAC 0x27 0x54 0x00 0x01 0xDC 0x02
05: 0xAB 0xC6 0x6B 0x0F 0xAC 0x27 0x6E 0x67 0x19 0xDC 0x08
06: 0xAB 0xC6 0x6B 0x11 0xAC 0x27 0x23 0x6E 0x67 0xDC 0x26
07: 0xAB 0xC6 0x6B 0x12 0xAC 0x27 0x84 0x04 0x02 0xDC 0x01
08: 0xAB 0xC6 0x6B 0x13 0xAC 0x27 0x8E 0x41 0x51 0xDC 0x22
09: 0xAB 0xC6 0x6B 0x14 0xAC 0x27 0x98 0x55 0xB1 0xDC 0x2D
10: 0xAB 0xC6 0x6B 0x15 0xAC 0x27 0xA2 0xC1 0x88 0xDC 0xC3
11: 0xAB 0xC6 0x6B 0x16 0xAC 0x27 0xAC 0x2B 0x57 0xDC 0x99
12: 0xAB 0xC6 0x6B 0x17 0xAC 0x27 0xB5 0x6B 0xFC 0xDC 0x04
13: 0xAB 0xC6 0x6B 0x18 0xAC 0x27 0xBF 0x17 0x07 0xDC 0xBE
```

Figure 4 The sequences set processed by our probability multiple sequence alignment

The experimental results and analysis of format inference.

Using the communication data encapsulated by Ethernet-IP-ICMP protocol as input data, the sequences set processed by our probability multiple sequence alignment without prior knowledge is shown in Fig.5

```
__ __ __ __ 7F B4 0C 74 8F 86 AD B7 08 00 45 00
__ __ __ __ 7F B4 0C 74 8F 86 AD B7 08 00 45 00
__ __ __ __ 7F B4 0C 74 8F 86 AD B7 08 00 45 00
__ __ __ __ 7F B4 0C 74 8F 86 AD B7 08 00 45 00
__ __ __ __ 7F B4 0C 74 8F 86 AD B7 08 00 45 00
__ __ __ __ 7F B4 0C 74 8F 86 AD B7 08 00 45 00
8F 86 AD B7 7F B6 0C 7C __ __ __ __ 2C 00 45 00
8F 86 A5 B7 7F B4 0C 74 __ __ __ __ 08 00 45 00
8D 86 AD B7 7F B4 0C 74 __ __ __ __ 08 02 45 20
8F 84 AD B7 5F B4 0C 74 __ __ __ __ 08 00 45 00
8F 86 AD B7 7F B4 04 74 __ __ __ __ 08 20 05 00
8F 86 AD B7 7F B4 0C 74 __ __ __ __ 08 00 45 00
```

Figure 5 The sequences set processed by probability multiple sequence alignment

After the calculation of various statistics, such as change rate, mean, variance and ratio of major, etc, is completed. We can found that the construction of the first 8 characters in the frames shows a cross feature. Because the change rates of column 1 to column 4 are almost equal, and column 5 to column 8, column 9 to column 12 have the same characteristic, so this 12 columns could be divided into 3 domain fields respectively. And the change rate of column 4 and 5, column 8 and 9 are different obviously, so they are the boundary of different domain fields. On the other hand, the ratio of major character of the front four columns close to 50%, and the middle four columns close to 100%, and the last four columns close to 50% also. We conclude that the fields with the cross characteristic maybe the address fields of both sides of communication.

The complete reverse result of protocol format is shown in Table 1. We can see from Table 1 that the algorithm can parse the frame format of the unknown protocol in byte level. The value of some adjacent fields in the captured packets is not changed all the time, which leads to these fields are merged into a larger constant field by the algorithm. For this case, we can increase the amount of data collected to highlight the gap of characteristics between the adjacent fields, and separate these fields into different domain fields.

Table 1 Format identified by format reverse analysis VS format defined in protocol specification

Format defined in protocol specification			Format identified by format reverse analysis without prior knowledge	
Ethernet protocol header	6 byte	Destination MAC address	6 byte	Cross field
	6 byte	Source MAC address	6 byte	
	2 byte	Type	6 byte	Constant field
IP protocol header	4 bit	Version		
	4 bit	Header length		
	1 byte	Type of service		
	2 byte	Total length		
	2 byte	Identifier	2 byte	Gradient field
	2 byte	Fragmented offset	2 byte	Constant field
	1 byte	Time to live	1 byte	Unknown field
	1 byte	Protocol	1 byte	Constant field
	2 byte	Header checksum	2 byte	Check field
	4 byte	Source IP address	4 byte	Cross field
	4 byte	Destination IP address	4 byte	
ICMP protocol header	1 byte	Type	1 byte	Unknown field
	1 byte	Code	1 byte	Constant field
	2 byte	Checksum	2 byte	Check field
	2 byte	Identifier	2 byte	Constant field
	2 byte	Serial number	2 byte	Gradient field

Summary

Protocol format is an important part of the protocol specification, through which the grammar logic of the protocol could be understood. To solve the format reverse problem of unknown

communication protocol, we designed a format reverse analysis method for unknown binary protocol from the communication data which based on alignment of fields and differential analysis of statistics. We designed the probability alignment algorithm for binary protocol frames by joining HMM analysis method into the basis of traditional progressive multiple sequence alignment. The corresponding fields in the frames could be aligned accurately through the probability alignment, and then various statistics of fields are calculated, and finally the boundary of adjacent fields in the frame and the format of protocol are identified according to the different characters of various statistics. In order to verify the effectiveness of the proposed method, the Internet protocol data is used as test data. The experimental results show that the probability alignment can effectively eliminate the negative impact of special frames and improve the accuracy of multiple sequence alignment results. And the protocol format reverse analysis method can effectively identify the format specification of binary communication protocol.

Acknowledgment

This work is supported by China Academy of Engineering Physics (CAEP) Project 2012A0403021, 2014A0403020, and by Laboratory of Network Security and Trusted Software of CAEP Project J-2014-zd-03.

References

[1] Z. Zhang, Q. Y. Wen, W. Tang. Sruvey of mining protocol specifications. Computer Engineering and Applications[J], 2013,49(9):1-9.(in chinese)

[2] M. Li, S. Z. Yu. Noise-Tolerant and optimal Segmentation of Message Formats for Unknown Application Layer Protocols[J]. Journal of Software, 2013,24(3):604-617.(in chinese)

[3] H. Dreger, A. Feldmann, M. Mai, et al. Dynamic application layer protocol analysis for network intrusion detection [C]//USENIX Security Symposium, Vancouver, Canada, 2006:257-272.

[4] W. M. Li, A. F. Zhang, J. C. Liu, et al. An Automatic Network Protocol Fuzz Testing and Vulnerability Discovering Method[J]. Chinese Journal of Computers, 2011, 34(2): 242-255.

[5] W. Cui, J. Kannan, H. Wang. Discoverer: Automatic protocol reverse engineering from network traces[C]//16th USENIX Security Symposium, USENIX, 2007: 199-212.

[6] Antunes João, Neves Nuno, Paulo Verissimo. Reverse engineering of protocols from network traces. [C]// 18th Working Conference on Reverse Engineering, WCRE, 2011: 169-178.

[7] Y. P. Wang, X. C. Yun, M. Z. Shafiq, et al. A semantics aware approach to automated reverse engineering unknown protocols[C]// 20th IEEE International Conference on Network Protocols (ICNP), 2012: 1-10

[8] M. Beddoe. Protocol information project [EB/OL].[2012-2-18]. http://www.4tphi.net/~awalters/PI/pi.pdf.

[9] F. Pan, Z. Hong, Y. X. Du, et al. Recursive Clustering Based Method for Message Structure Extraction[J]. Journal of Sichuan University, 2012,44(6):137-142. (in chinese)

[10] J. CABALLERO, D. SONG. Automatic protocol reverse-engineering: message format extraction and field semantics inference[J].Computer Networks， 2013,57(2) : 451-474

[11] F. PAN, Z. HONG, Y. DU, et al. Efficient protocol reverse method based on network trace analysis. [J]. International Journal of Digital Content Technology and its Applications, 2012, 20(6) : 201- 210.