IET Communications

Research Article

ISSN 1751-8628
Received on 17th August 2015
Revised on 25th November 2015
Accepted on 23rd December 2015
doi: 10.1049/iet-com.2015.0797
www.ietdl.org

Bit-oriented format extraction approach for automatic binary protocol reverse engineering

Siyu Tao ✉, *Hongyi Yu*, *Qing Li*

National Digital Switching System Engineering and Technological Research Center (NDSC), 450001, Zhengzhou, People's Republic of China
✉ *E-mail: peachforworking@foxmail.com*

Abstract: Protocol message format extraction is a principal process of automatic network protocol reverse engineering when target protocol specifications are not available. However, binary protocol reverse engineering has been a new challenge in recent years for approaches that traditionally have dealt with text-based protocols rather than binary protocols. In this study, the authors propose a novel approach called PRE-Bin that automatically extracts binary-type fields of binary protocols based on fine-grained bits. First, a silhouette coefficient is introduced into the hierarchical clustering to confirm the optimal clustering number of binary frames. Second, a modified multiple sequence alignment algorithm, in which the matching process and back-tracing rules are redesigned, is also proposed to analyse binary field features. Finally, a Bayes decision model is invoked to describe field features and determine bit-oriented field boundaries. The maximum a posteriori criterion is leveraged to complete an optimal protocol format estimation of binary field boundaries. The authors implemented a prototype system of PRE-Bin to infer the specification of binary protocols from actual traffic traces. Experimental results indicate that PRE-Bin effectively extracts binary fields and outperforms the existing algorithms.

1 Introduction

Automatic protocol reverse engineering [1] processes undocumented protocols to deduce message formats without a priori knowledge of protocol specifications. With the help of closed-protocol analysis, network protocol reverse engineering (NPRE) [2] plays an important role in network management and security applications (e.g. intrusion detection system [3] and vulnerability mining [4]). To date, network-based, program-based and hybrid methods have constituted the types of NPRE techniques [5].

However, automatic binary protocol reverse engineering (BPRE) based on network traces [1, 5] is a new and difficult process. According to [6], more than 40% of traffic on backbone networks worldwide comprises protocols of non-public descriptions within a quite considerable proportion of binary protocols for many controllers (e.g. C&C botnet servers [7], data link networks [8] and wireless networks [8]). The NPRE information available for binary protocols is in bit streams which contain less priori knowledge than text-based protocols. Meanwhile, the existing study [9] on closed-protocol reverse critically depended on 12-year manual labours when binary protocol specifications were private; such efforts are time consuming and error prone.

An early attempt in automatic NPRE, the Protocol Informatics project [10], applied a multiple sequence alignment (MSA) algorithm [10] to extract the protocol structure and infer message fields from network traces. Another proposed approach, Discoverer [11], extracted message formats from sequences of protocols, using a foremost tokenisation process by exploiting delimiters or character encodings. PI project and Discoverer, regarded as classic projects that are based on network traces in NPRE, are not applicable to BPRE for two reasons: (i) traditional analysis models take a byte as the minimum unit whereas fields of binary protocols often consist of only several bits; (ii) token delimiting depends on printable characters or universal character sets whereas binary protocols are virtually transparent within closed encoding or no strings. For these reasons, similar proposed methods [12, 13] are also unable to effectively reverse engineer binary protocols.

Various studies for BPRE have been explored in recent years. For example, variance of the distribution of variances (VDV) [2] selects binary features with a special focus on binary protocol format extraction based on network traces. A hidden semi-Markov model [14] for the segmentation of unknown application-layer protocols was proposed. A practical system called AutoReEngine [15] based on the Apriori algorithm and positions deduced protocol keywords and formats. An unsupervised approach [16] based on information entropy was depicted to extract protocol feature keywords from traffic traces. While these methods may be theoretically suitable for reverse engineering of binary protocols, the effectiveness has not been evaluated experimentally and maturely on datasets of binary frames.

This paper presents an approach for a new issue of automatic BPRE based on network traces. It addresses two main challenges: (i) binary protocols rarely contain human-readable strings; also, binary field boundaries in general cannot be identified by common delimiters such as whitespace characters when they are reverse engineered; (ii) fields of binary protocols are flexible on the length attribute, which usually consist of several bits, not bytes. Therefore, we implemented our approach, called PRE-Bin, on a prototype system that extracts fields of binary frames from binary protocols. In the proposed system, the silhouette coefficient is used to reduce the time and space complexity of confirmed optimal clusters and enhance the MSA algorithm to exploit back-tracing rules in order to derive gap features from clusters. We have also designed a bit-oriented probability decision criterion based on the Bayes model to determine the binary field boundaries and use the maximum a posteriori (MAP) criterion to optimally estimate the format of the binary frames. The results of simulation experiments, in which the proposed system was used to analyse automatic identification system (AIS) protocol, high-level data link control (HDLC) protocol, network basic input/output system (NetBIOS) protocol and internet control message protocol (ICMP), indicate that its coverage in discovering actual fields, its correctness in associating actual and inferred fields, and its closeness of actual and inferred bit positions are at least 75%, 75%, and 85%, respectively. Compared with existing algorithms that take binary network traces as input (e.g. PI project, Discoverer, VDV and AutoReEngine), our approach can obtain higher accuracy in extracting binary fields under difficult conditions such as there is less priori experience with binary frames and binary fields have flexible boundaries.

The remainder of this paper is organised as follows. Section 2 defines the problem and definition of binary protocols. Section 3 describes the design and elaborates the details of PRE-Bin. Section 4 demonstrates PRE-Bin's ability to infer binary frames via simulation experiments and evaluates the proposed approach. Finally, Section 5 concludes this paper.

2 Problem statement

2.1 Binary protocol

This paper discusses a new issue, BPRE, for reasons that traditional NPRE methods deal with character-oriented protocols and current methods have not been clearly demonstrated to be effective on binary frames. Our novel method catering to BPRE based on network traces is intended to resolve binary field extractions. The concepts of character-oriented and bit-oriented protocols in the link layer were originally described in [17]. We enrich these concepts in NPRE as follows:

- *Character-oriented protocol*: The fields often consist of several bytes. The packets may be printable, usually using open character encodings or delimiters.
- *Bit-oriented protocol*: The fields often consist of several bits. The frames rarely use readable character encodings and reflect the transparent property of bit streams.

However, differences between the concepts of text and binary protocols are more complex and ambiguous. Some acknowledged binary protocols contain byte-defined fields such as Bit Torrent (BT) protocol and Server Message Block (SMB) protocol. The paper considers that the core standard to distinguish text-based and binary protocols is that binary protocols have no distinguishable attributes on binary-type fields [14] for the posterior observation as follows.

- *Flexible length*: The length attribute of fields may appear not only byte-defined but also bit-defined situations among different binary protocols.
- *Abundant value*: The fields may avoid the open encoding set or use closed encoding set so that whichever gets reversed on delimiters will be tough or blind.

2.2 Problem definition

The technical feasibility analysis of NPRE given in [1] can be summarised in terms of the locality principle of protocols, which states that continuous traffic within an appropriate space and time may contain similar formats, close semantic types, and order-constrained messages. According to Li and Yu [14], protocol fields can be grouped into four main structural types: type-length-value (TLV) types, ASCII types, binary types and indicator types. Fields of binary protocols, for example the SMB protocol, AIS protocol and HDLC protocol, tend to have attributes of the binary value and fixed length. In this paper, the structure of binary protocols belongs to binary type typically as opposed to TLV type that often appears in text-based protocols as shown in Fig. 1.

The fields of a binary frame are denoted as $\overline{f_1 f_2 \ldots f_a}$, each field is labelled as $\overline{f_{i+1}} = \overline{x_{\text{inter}(i)+1} \cdots x_{\text{inter}(i+1)}}, (i \geq 0)$ in which $x \in \{\backslash 0\text{b}0, \backslash 0\text{b}1\}$ and the field boundaries *inter*(*i*) and *intra*(*i*) are confirmed by the protocol specification. The optimal destination, which is difficult for protocol reverse engineering, is to resolve the joint probability distribution defined as $P(x_r, \text{inter}, \text{intra})$ where r is the bit position. If $\theta = \text{inter}(c)$, $r \in [1, \theta]$ and *intra*(*i*) is transparent and unreachable in binary protocols, this paper attempts to resolve a posteriori probability problem defined as $P(r = \text{inter} | x_r, \ r \in [1, \theta])$ for BPRE.

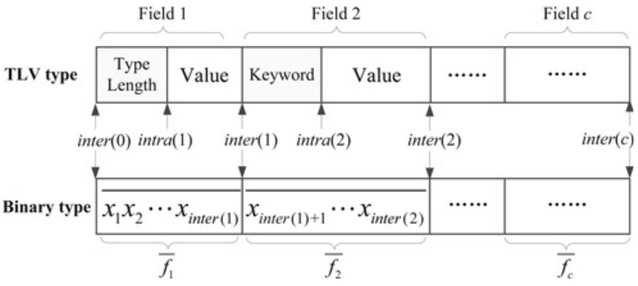


Fig. 1 *Type-length-value (TLV) and binary types*

However, we cannot obtain any information about closed protocols (e.g. keywords of fields and delimiters) or do not exploit any boundaries (e.g. *inter*(*i*) and *intra*(*i*)) when reversing binary protocols. The flexible length and abundant value without delimiters may easily cause the ambiguity of field boundaries, which traditional methods ignore. A new extraction feature is needed to pull probable positions related to boundaries, and a decision model should be invoked to determine which boundary is the most likely field endpoint.

3 Design

3.1 Overview

In this paper, the clustering algorithm is used to classify the close frames among different formats and the MSA algorithm is used not only to compute the similarity relation among messages but also to excavate the structure of sequences. Most importantly, a Bayes model is designed for redundancy information processing and optimal field extraction. The details of our proposed PRE-Bin for binary protocol analysis are shown in Fig. 2.

- *Iterative clustering algorithm:* Hierarchical clustering [10] provides a more comprehensive classification than the *k*-means algorithm but has the difficulty to determine the optimal clusters. To reduce its time-space complexity, we introduce the silhouette coefficient [18] to guide the optimal number of clusters. Subsequently, we import hierarchical clustering based on the silhouette coefficient to turn the sample space of the similar message format into subspaces. Then, we utilise iterative clustering to produce various subspace sizes, up to the optimal number, and offer a diversity of features for improved MSA algorithm.
- *Improved MSA algorithm:* The MSA algorithm from PI project cannot be applied to BPRE directly for the reason that gap insertions in PI project cause to slide gap features of binary protocols. To rectify the flaws in the original MSA algorithm and conquer feature extraction of boundaries in binary fields, we propose an improved MSA using bit-oriented matching and set it to operate on a sequential alignment. Further, to avoid a gap shift

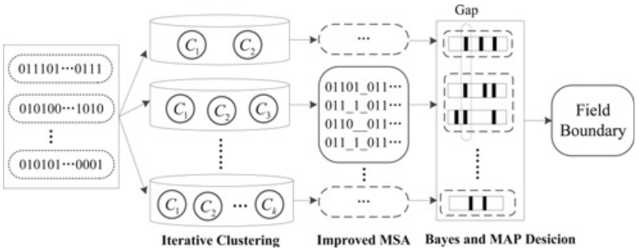


Fig. 2 *Design of PRE-Bin*

from the same field, we design a replacing rule that displaces the gap insertion in the back-tracing rule to provide the uniform output of the pairwise alignment for the bit-oriented alignment.

- *Decision fusion algorithm:* In the subspace cluster, we leverage the locality principle of protocols to present a bit-oriented decision criterion with frequency distribution of gaps to determine the boundaries of binary fields. In spite that determining the joint probability distribution of binary fields is difficult, we consider that extracted features from the posterior distribution of gaps can be described by the Bayes model to locate the potential field boundaries. Among clusters of the subspace, the fusion of inferred formats is formed using the Bayes model and MAP criterion for optimal estimation.

3.2 Iterative hierarchical clustering based on silhouette coefficient

The iterative hierarchical clustering based on silhouette coefficient is exploited to keep the precision and reduce the time-space complexity of the hierarchical clustering. Each subspace includes different clustering sizes to provide the diversity of frames for next algorithm called improved MSA. We assume that the input of PRE-Bin is the sample space D consisting of binary frames from the similar message format and D will be divided into different subspaces defined as *SubD*.

Step 1: The sample space D uses k-means algorithm with the selection of a random centroid and a clustering number k. A subspace with clusters is output as $\mathrm{KC}(k) = \{\mathrm{KC}_1, \ldots, \mathrm{KC}_k\}$.
Step 2: When $k = 2$ and $\gamma \in [1, k]$, the silhouette coefficient s_i of an object i is defined as in (1).

$$s_i = \sum_{i=1}^{\left|\mathrm{KC}_\gamma\right|} \frac{\beta_i - \alpha_i}{\max(\alpha_i, \beta_i)} \tag{1}$$

where α_i and β_i are shown in as (2) and (3).

$$\alpha_i = \frac{\sum_{i,j \in \mathrm{KC}_\gamma, \mathrm{KC}_\gamma \in \mathrm{KC}(k)} \mathrm{distance}(i, j)}{\left|\mathrm{KC}_\gamma\right|} \tag{2}$$

$$\beta_i = \min_{\mathrm{KC}_\gamma \neq \mathrm{KC}_\eta, \mathrm{KC}_\eta \in \mathrm{KC}(k)} \left(\frac{\sum_{i \in \mathrm{KC}_\gamma, j \in \mathrm{KC}_\eta} \mathrm{distance}(i, j)}{\left|\mathrm{KC}_\eta\right|} \right) \tag{3}$$

Step 3: For all n objects in $KC(k)$, silhouette coefficient $s_{\mathrm{KC}(k)}$ of $KC(k)$ is defined as in (4) for the subspace.

$$s_{\mathrm{KC}(k)} = \frac{1}{n} \sum_{i=1}^{n} s_i \tag{4}$$

Step 4: Compute $s_{\mathrm{KC}(k)}$ repeatedly by $\mathrm{threshold}_1$ times and the average of silhouette coefficient $\overline{s_{\mathrm{KC}(k)}}$ is defined as in (5) for the subspace $KC(k)$.

$$\overline{s_{\mathrm{KC}(k)}} = \frac{1}{\mathrm{threshold}_1} \sum_{i=1}^{\mathrm{threshold}_1} s_{\mathrm{KC}(k)} \tag{5}$$

Step 5: Increase k one by one while recording each $\overline{s_{\mathrm{KC}(k)}}$ of subspaces $KC(k)$ and go to step 1 until $k \in [2, \mathrm{threshold}_2]$ is not satisfied. Search the maximum $\overline{s_{\mathrm{KC}(k)}}$ to denote k as k_{opt}.
Step 6: Set l and *stepsize*. The number of output clusters l is defined as (6). Unweighted pair-group method with arithmetic means (UPGMA) [10] is utilised to generate clusters of subspace defined as (7).

$$l \in \left\{ l \mid l \in [2, k_{\mathrm{opt}}] \text{ and } (l\%\mathrm{stepsize}) = 0 \right\} \tag{6}$$

$$\mathrm{Sub}D = \left\{ \mathrm{HC}_1, \ldots, \mathrm{HC}_l, \ldots, \mathrm{HC}_{k_{\mathrm{opt}}} \right\} \tag{7}$$

3.3 Improved MSA based on replacing rules

Needleman–Wunsch algorithm quoted in PI project is altered in this paper to form our improved MSA algorithm. Each cluster HC_l needs to pass through improved MSA based on revised rules and will be affected by our improved details as follows.

- *Initialisation:* The pairwise sequence alignment of PI project is out-of-order because the output of pairwise sequences is two sequences with gaps and independent on the order of a pair. We turn it into sequential matching by appointing one of pairwise sequence as a template sequence and the other as a matching sequence.
- *Similarity scoring:* The score of match and penalty should be distinct. Actually the match score we set is 1, mismatch score is –1 and gap penalty score is –2.
- *Back-tracing rules:* Back-tracing rules of PI project have a gap insertion that a gap is inserted into a sequence and there is a gap between two symbols in a sequence. We use replacement operation instead of gap insertion and gap replacement covers the present symbol.
- *Output:* When triggering gap insertions into a matching sequence, gap replacement will be taken over. When triggering gap insertions into a template sequence, gap insertion is stopped. Thus the output will be one sequence with gaps.

Then, multiple sequence alignment performs order-pair comparisons in each cluster of each subspace and improved MSA algorithm combines the global and local alignment to output one sequence representing the similarity between a matching and a template sequence as shown in Algorithm 1 (see Fig. 3), which uses Python expression in some statements.

3.4 Decision fusion based on Bayes and MAP criterion

Through improved MSA, each cluster will generate a lot of sequences with gaps. The analysis on sequences with gaps encounters two questions about determining boundaries of binary fields and combining inferred formats. Details are explained as follows.

3.4.1 Bit-oriented decision fusion for intra-clusters based on Bayes decision rules: We suppose that a subspace denoted as $\mathrm{Sub}D = \{\mathrm{HC}_1, \ldots, \mathrm{HC}_l, \ldots, \mathrm{HC}_{k_{\mathrm{opt}}}\}$ has been reformed into a new subspace denoted as $\mathrm{Sub}D' = \{\mathrm{HCG}_1, \ldots, \mathrm{HCG}_l, \ldots, \mathrm{HCG}_{k_{\mathrm{opt}}}\}$ by improved MSA. A sequence $\mathrm{Sequence}_m$ in HCG_l is defined as in (8).

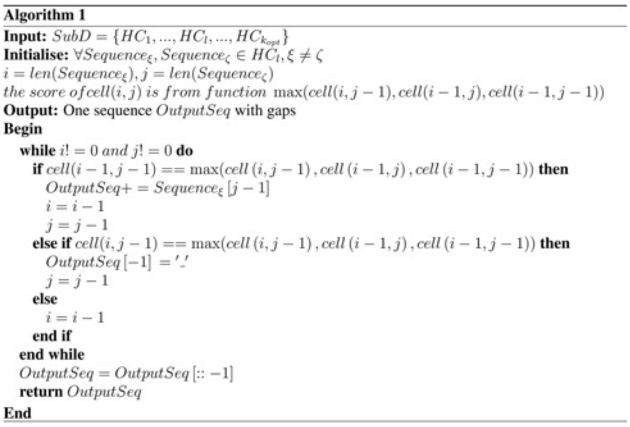
$$\mathrm{Sequence}_m = \overline{x_1 x_2 \ldots x_r \ldots x_\theta} \tag{8}$$

x_r, the value of r_{th} bit position, has priori and target attributes as in (9) and (10).

$$u_r \in \left\{ \begin{pmatrix} Y \\ N \end{pmatrix} | x_r \begin{pmatrix} \text{is} \\ \text{is not} \end{pmatrix} \text{field boundary} \right\} \tag{9}$$

$$\omega_r \in \left\{ \begin{pmatrix} Y \\ N \end{pmatrix} | x_r \begin{pmatrix} \text{has a} \\ \text{has no} \end{pmatrix} \text{gap} \right\} \tag{10}$$

Bayes decision rules will be exploited to analyse whether a bit position declared as a gap is the endpoint of a field or not. Equation (11) is defined to bring the same bit position into a longitudinal vector from sequences in a cluster of a subspace, in which $x_r^{\mathrm{Seq}_m}$ represents



```
Algorithm 1
Input: $SubD = \{HC_1, ..., HC_l, ..., HC_{k_{opt}}\}$
Initialise: $\forall Sequence_\xi, Sequence_\zeta \in HC_l, \xi \neq \zeta$
$i = len(Sequence_\xi), j = len(Sequence_\zeta)$
$the\ score\ of cell(i,j)\ is\ from\ function\ \max(cell(i,j-1), cell(i-1,j), cell(i-1,j-1))$
Output: One sequence $OutputSeq$ with gaps
Begin
  while $i! = 0\ and\ j! = 0$ do
    if $cell(i-1,j-1) == \max(cell\,(i,j-1)\,, cell\,(i-1,j)\,, cell\,(i-1,j-1))$ then
      $OutputSeq+ = Sequence_\xi\,[j-1]$
      $i = i - 1$
      $j = j - 1$
    else if $cell(i,j-1) == \max(cell\,(i,j-1)\,, cell\,(i-1,j)\,, cell\,(i-1,j-1))$ then
      $OutputSeq\,[-1] = '\_'$
      $j = j - 1$
    else
      $i = i - 1$
    end if
  end while
  $OutputSeq = OutputSeq\,[:: -1]$
  return $OutputSeq$
End
```

Fig. 3 *Improved back-tracing process of Neddlema-Wunsch algorithm*

the value on the r_{th} bit position of the Sequence$_m$.

$$\overleftarrow{\boldsymbol{t}_r} = \langle x_r^{\mathrm{Seq}_1}, x_r^{\mathrm{Seq}_2}, \ldots, x_r^{\mathrm{Seq}_m}, \ldots \rangle, \{\mathrm{Seq}_1, \ldots, \mathrm{Seq}_m, \ldots\} \in \mathrm{HCG}_l \tag{11}$$

According to Bayes decision theory, the classification model can be defined as in (12) and (13) with the decision function defined as in (14).

$$p_r = P(u_r = Y | \omega_r = Y) \tag{12}$$

$$q_r = P(u_r = Y | \omega_r = N) \tag{13}$$

$$g(x_r) = \sum_{m=1}^{|\mathrm{HCG}_l|} \left[x_r^{\mathrm{Seq}_m} \cdot \ln \frac{p_r}{q_r} + (1 - x_r^{\mathrm{Seq}_m}) \cdot \ln \frac{1 - p_r}{1 - q_r} \right] + \ln \frac{P(u_r = Y)}{P(u_r = N)} \tag{14}$$

However, it is hard to calculate p_r, q_r and $P(u_r = Y)$ for protocol reversing lacking of priori information, two proposals are shown as follows: (i) choose appropriate disclosure binary protocols close to characteristics of undocumented protocols for acquiring training labels; (ii) the empirical law for approximate values may substitute variables. In PRE-Bin, (15) and (16) were proposed as the empirical law, indicating that the probability of gaps on a bit position is displaced by the frequency of gaps. It is important to note that the *offset* can be added to avoid the sparse and illegal value.

$$P(\omega_r = Y) = \frac{\mathrm{Num}(\omega_r = \text{gap in } \overrightarrow{\boldsymbol{t}_r}) + \text{offset}}{\left|\overrightarrow{\boldsymbol{t}_r}\right|} \tag{15}$$

$$P(\omega_r = N) = \frac{\mathrm{Num}(\omega_r = 0 \text{ or } 1 \text{ in } \overrightarrow{\boldsymbol{t}_r}) - \text{offset}}{\left|\overrightarrow{\boldsymbol{t}_r}\right|} \tag{16}$$

To extract bit-oriented boundaries from bit positions in sequences with gaps, we summarise that it is dependent not only on distributions of gaps in HCG$_l$ but also on gradient distributions of gap distributions. Equation (17) shows that the decision criterion focuses on gap distribution by Bayes decision rules while the other decision rule was established in (18) to dissect incremental changes of the gap distribution. Both (17) and (18) reflect that the more probability of gaps on a bit position, the more possibility of a bit position attached to field endpoints.

Table 1 Dataset

Protocols	Message format	Number
AIS	type 1	1000
	type 3	1000
	type 4	1000
	type 18	1000
	type 24b	500
HDLC	header + payload	1000
NetBIOS	datagram	1000
ICMP	type 0	500
	type 8	500

$$\underset{r}{\operatorname{argmax}}\left[g(x_r) \geq \mathrm{threshold}_3\right] \tag{17}$$

$$\underset{r,r+1}{\operatorname{argmax}}\left[\left|P(\omega_r = Y) - P(\omega_{r+1} = Y)\right| \geq \mathrm{threshold}_4\right] \tag{18}$$

3.4.2 Format fusion for inter-clusters based on MAP criterion: If $\mathrm{HCG}_l \in \mathrm{Sub}D'$ generates a group of inferred bit positions denoted as $\overrightarrow{\mathrm{Inferred}_{\mathrm{HCG}_l}}$ which represents effective positions r_i passed through the probability or frequency decisions of gaps by methods above, inferred results of other clusters should be merged by supplementing distinct positions and combining the same position in a cluster with probability or frequency of gaps. For merged results from clusters of the subspace, (19) can be leveraged to draw the optimal estimation of the subspace. (see (19))

$$\underset{r}{\operatorname{argmax}}\left\{P(r_j) > \mathrm{threshold}_5 \middle| \forall a \in \left(\bigcup_{r_j} \underset{l \in [1, k_{\mathrm{opt}}]}{\forall} \overrightarrow{\mathrm{Inferred}_{\mathrm{HCG}_l}} \in \mathrm{Sub}D'\right)\right\} \tag{19}$$

4 Simulation

4.1 Conditions

On the basis of actual traffics of binary protocols, including the AIS, HDLC, NetBIOS and ICMP protocols, the simulation selects representative samples as listed in Table 1.

Detailed descriptions of the AIS, HDLC, NetBIOS and ICMP protocols are provided in [19–22] which characterise binary-type fields. Several parameters can be set as (20)–(25) where *max* and *min* mean that the maximum and minimum values of the gap distribution, respectively.

$$\mathrm{threshold}_1 = 10 \tag{20}$$

$$\mathrm{stepsize} = 1 \tag{21}$$

$$\mathrm{threshold}_2 = 30 \tag{22}$$

$$\mathrm{threshold}_3 = 1.0 \tag{23}$$

$$\mathrm{threshold}_4 = (\max - \min) \times 35\% \tag{24}$$

$$\mathrm{threshold}_5 = (\max - \min) \times 50\% \tag{25}$$

The inferred positions are denoted as r_j with the true positions of the field specification denoted as *inter*(i). The operations of the single intersection are defined as (26) and (27) which mean the moduli about the difference of positions and the overlapping of fragments, respectively. Equation (28) shows the operation of the intersection with sets. According to (26)–(28), the metrics [11] shown in (29)–(31) for a comprehensive assessment are revised indicators with more practicability for evaluating BPRE.

$$\mathrm{position}_i \cap \mathrm{position}_j = \left|\mathrm{position}_i - \mathrm{position}_j\right| \tag{26}$$

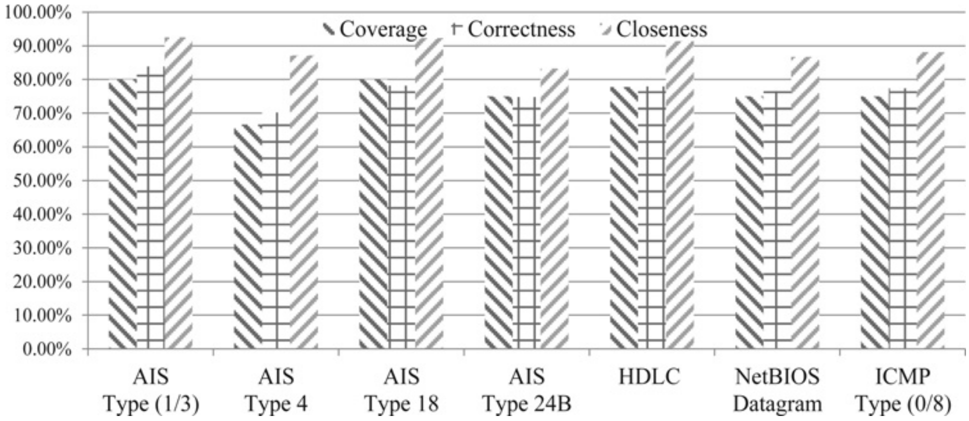


Fig. 4 *Coverage, correctness and closeness of PRE-Bin*

$$\overline{x_{r_1}\ldots x_{r_i}\ldots x_{r_j}} \cap \overline{x_{r_i}\ldots x_{r_j}\ldots x_\theta} = \left|\overline{x_{r_i}\ldots x_{r_j}}\right| \tag{27}$$

$$\{\text{set}_1\} \cap \{\text{set}_2\} = \{\forall i \in \{\text{set}_1\} \cap \forall j \in \{\text{set}_2\}\} \tag{28}$$

- *Coverage:* The coverage of inferred fields as shown in (29) indicates the consistency of inferred and actual fields under some uncertainty, such as allowing the 2-bit offset error.

$$\frac{\left|\bigcup_{\left(\{r_j\}\cap\{\text{inter}(i)\}\right)\le\text{allowance error}} \{r_j\}\right|}{\left|\{\text{inter}(i)\}\right|} \times 100\% \tag{29}$$

- *Correctness:* The correctness of the inferred fragment as shown in (30) describes the similarity between the inferred fragment and the actual field.

$$\frac{2\times\max\left(\left\{\overline{x_{r_j}\ldots x_{r_{j+1}}}\right\}\cap\left\{\overline{x_{\text{inter}(i)}\ldots x_{\text{inter}(i)+1}}\right\}\right)}{\left|\overline{x_{r_j}\ldots x_{r_{j+1}}}\right| + \left|\overline{x_{\text{inter}(i)}\ldots x_{\text{inter}(i)+1}}\right|} \times 100\% \tag{30}$$

- *Closeness:* The closeness of inferred positions as shown in (31) reflects the proximity of inferred positions and the actual field boundary.

$$\left(1-\frac{\sum_{\forall r_j\in\overline{x_{\text{inter}(i)}\cdots x_{\text{inter}(i)+1}}} \max\left(r_j\cap\text{inter}(i),\, r_j\cap\text{inter}(i+1)\right)}{\left|\{r_j\}\right|}\right) \times 100\% \tag{31}$$

4.2 Evaluations

According to Table 1 and (29)–(31), Fig. 4 which indicates the evaluation of the dataset can be obtained. The result shows that the averages of coverage and correctness are at least 75% while the average of closeness is greater than 85%, and all protocols in the dataset possessing binary-type fields gain the binary field extraction well by PRE-Bin. In addition, it is important to note that the inferred results of AIS message type 1 and type 3 are coincident, the inferred results of ICMP message type 0 and type 8 are coincident and the inferred result of HDLC displays a portion of the header specification in the picture with cutting off payloads.

4.3 Factors affecting PRE-Bin

For comprehensive analyses, we discuss three types of limitations, that is, dataset size, dataset diversity and parameter evaluation.

4.3.1 Dataset size: To evaluate the PRE-Bin with respect to the dataset size, we randomly sample from actual AIS message type 1 data as shown in Fig. 5. The parameters for the experiment are set as shown in (20)–(25) except for *stepsize* = 5.

Fig. 5 shows that when the size of dataset reaches 600 on AIS message type 1 data, the performance of PRE-Bin is stable and trustiness, which indicates that for these datasets of binary-type structures, PRE-Bin may acquire obvious effects.

4.3.2 Dataset diversity: The diversity of dataset may result in no frequency of gaps due to the constant fields in a continuous region out of a sequence. In a way, the diversity of dataset may result in the interference of wrong gaps due to dataset within the noise. First, we encounter the problem of no gaps appearing in AIS type 4 data and design a gap feature selection for searching more changes of gaps. We obtain Fig. 6 to show that after feature selecting process, the abundant results of gap judgment is satisfied until removing repeated frames or adding sequences under the unsupervised operation. The unsupervised operation is that when a region of the sequence does not appear the gaps in the gap distribution, we specifically collect the same position of that region among the datasets to search different values of regions as new added members into the datasets.

Second, to evaluate the PRE-Bin by considering the interference factor of binary frames, we choose AIS message type 1 data and make sufficient copies of one sequence to create datasets by repetitions of different ratios, as shown in Fig. 7, while the size of dataset being fixed on 1000. The parameters of the experiment are set as shown in (20)–(25) except for *stepsize* = 5. Fig. 6 reflects

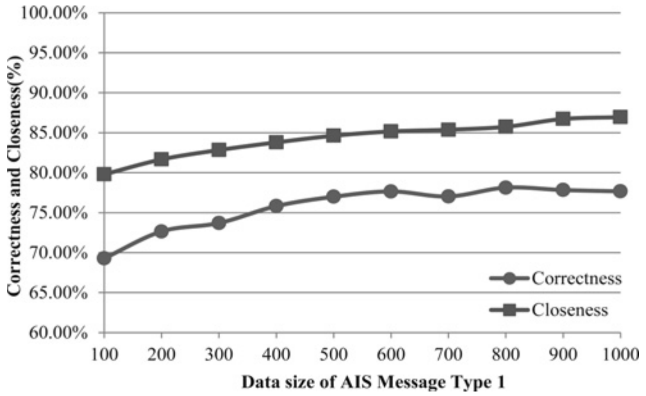


Fig. 5 *Different data size of AIS message type 1*

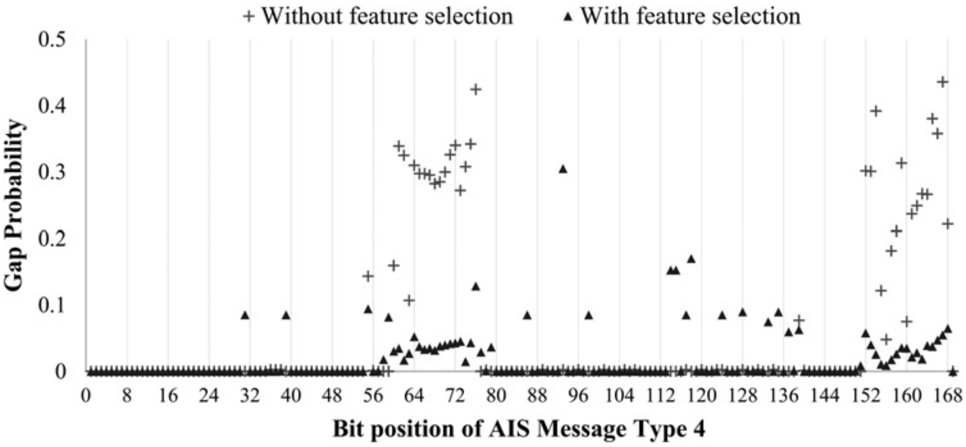


Fig. 6 *Gap probability distribution of AIS message type 4 after feature extraction*

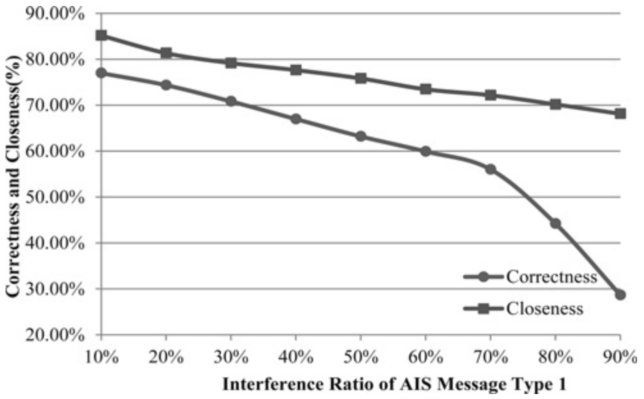


Fig. 7 *Different interference ratio of AIS message type 1*

that binary frames with constant fields may be treated as the interference factor in the dataset which is defined as the interference ratio of the repetitions to the sample space. Fig. 7 indicates that when the interference ratio goes beyond 30%, the correctness is less than 70% and the closeness is less than 80%. To ensure the quality of the PRE-Bin, the interference ratio threshold of tolerance is approximately 30%. When the interference ratio reaches 70%, the performance of PRE-Bin worsens; moreover, PRE-Bin cannot work when the interference ratio is close to 100%.

4.3.3 Parameter evaluation: The parameters of PRE-Bin involve the performance of the field boundary location. We evaluate the *stepsize* and the threshold of gap decisions as two aspects of important parameters. The control experiment of important parameters performs to follow two principles: to give contrast and to select single variable. The experiment takes AIS message type 1 data from Table 1 and set parameters according to (20)–(25) unless the single variable is claimed to renew.

- We set the *stepsize* as shown in Fig. 8 and keep other parameters unchanged. Fig. 8 shows that the *stepsize* affects the appearance of effective bit positions mapping to field boundaries. In the real situation, the HCG$_4$ generates the decision result which does not appear in other *HCG*s as well as the experiment of HDLC protocol shows this phenomenon so that the iterative process of *stepsize* may not omit this important position decisions. However, according to the practical situation, the operation of *stepsize* may bring two problems: (i) when the sample space *D* consists of frames which possess the accurate similarity, the activity of *stepsize* is weak which is evident in the experiment of high-refined frames without much changes; (ii) the time-space complexity of prototype system may be enhanced by increasing threshold$_2$ and decreasing *stepsize* while the decreasing *stepsize* may bring more precise results.
- We set the threshold$_4$ as shown in Fig. 9 and keep other parameters unchanged. Fig. 9 shows that the effective results will be reduced if the threshold of gap decisions increases and the redundant deduction of boundary decisions will be introduced if the threshold of gap decisions decreases.

4.4 Comparisons

To address the format analysis of binary protocols, we compare PRE-Bin with existing algorithms, which include PI project,

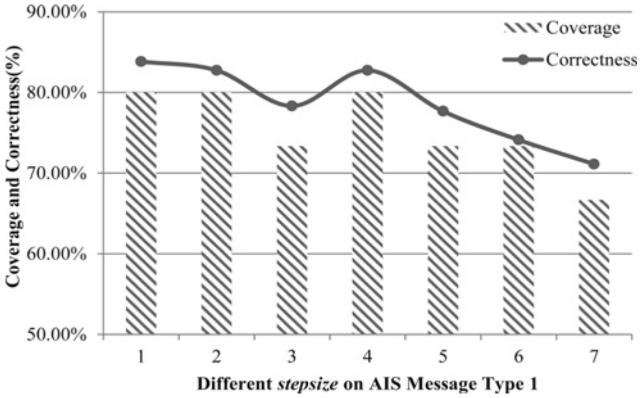


Fig. 8 *Different stepsize of AIS message type 1*

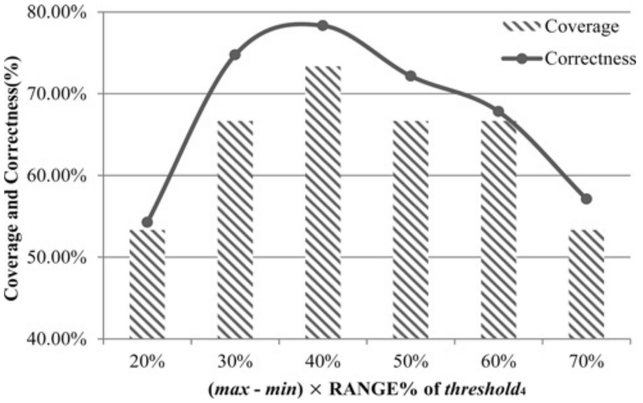


Fig. 9 *Range of threshold$_4$ on AIS message type 1*

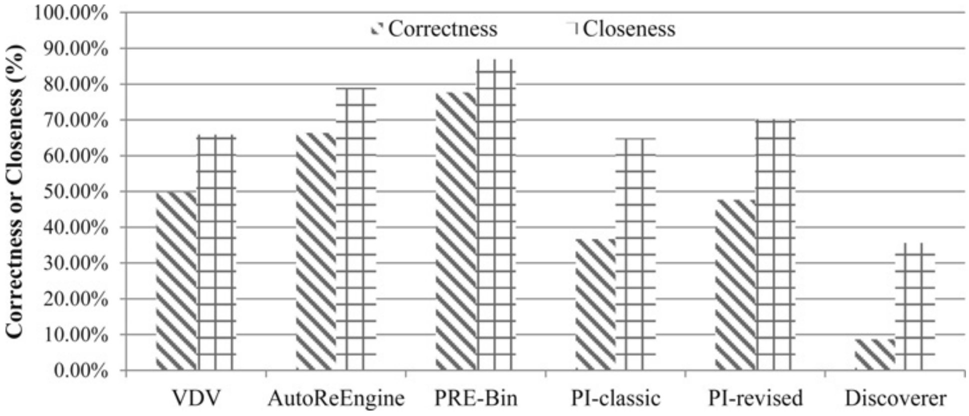


Fig. 10 *Correctness and Closeness of VDV, AutoReEngine, PRE-Bin, PI project and Discoverer on AIS message type 1*

Discoverer, VDV and AutoReEngine. First, we compare PRE-Bin with PI project and Discoverer to verify if these approaches can work on AIS binary protocol correctly or not. Then, we compare PRE-Bin with VDV and AutoReEngine to confirm whether they can work on AIS binary protocol effectively or not.

4.4.1 Comparing PRE-Bin with PI project and discoverer: Many binary sequences with gaps will be produced by PI project; to be precise, it is hard to identify field boundaries for no format reduction in PI project. In other words, the details of format extraction in PI project are not available for the reason that effective measures for field delimitation have not been proposed by PI project. Therefore, we generate the result of the format from the most similar cluster by PI project, called PI-classic, and append our format extraction algorithm into PI project, called PI-revised. In the experiment, we take the AIS message type 1 data and PRE-Bin parameters are set according to (20)–(25) except for *stepsize* = 5.

In this experiment as shown in Fig. 10, we verify our assumption that the classic algorithms such as PI project and Discoverer are not effective for BPRE. However, PI project with the MSA algorithm cannot be directly applied to BPRE for the following reasons:

- The demarcation of protocols in PI project is dependent on blocks of bytes and spaces.
- PI project has no decision algorithm to determine the protocol format in the final output.
- Gap insertions in PI project cause gap features of binary protocols to slide to the left or right.

Moreover, Discoverer represents a class of methods committed to application-layer or text-based protocol reverse engineering and performs incorrectly for binary frames. Fig. 10 shows that PI project without accurate field location gets the undesirability correctness and the accuracy of PRE-Bin is higher than that of PI project and Discoverer. The design of our approach was inspired by PI project but differed from PI project in two aspects. First, PI project has no format reduction so that PI-revised using our Bayes decision will be better than PI-classic. Second, PRE-Bin performs better than PI-revised on evaluations for that our design includes improved MSA and iterative clustering.

4.4.2 Comparing PRE-Bin with VDV and AutoReEngine: As shown in Fig. 10, VDV defines the byte as the basic feature unit to identify the most relevant fields in protocol sequences. Hence, it selects the binary field feature inaccurately. However, AutoReEngine leverages the Apriori algorithm to improve the accuracy of VDV based on positions without bytes limited and performs better than VDV for BPRE competently as show in Fig. 10. In this evaluation, our approach outperforms the current approaches, such as VDV and AutoReEngine, for two reasons. First, VDV can separate binary field boundaries in theory but get no disambiguation of boundary features as same as PI project. Second, PRE-Bin adopts an effective mechanism using the improved MSA and Bayes decision criterion, such that binary frames with boundary features are better identified by our model than a simple variance distribution or frequent-item model.

5 Conclusions

In this study, we designed a prototype system called PRE-Bin for binary field extraction of binary protocols. We consider the AIS protocol as a representative of protocols with cable formats, the HDLC protocol as a representative of link-layer protocols, while NetBIOS and ICMP protocols are considered to be the representatives of network-layer protocols. Testing PRE-Bin by binary frames in real situations demonstrates that our system achieves the target that binary-type fields of binary frames are extracted and performs better than existing approaches. In the meantime, we believe that both the dataset sampling theory and feature extraction of binary fields in PRE-Bin require further research. With the advent of the technical requirements about network protocol analysis, PRE-Bin provides new ideas for reversing binary protocols (e.g. gap distribution, Bayes model and MAP decision criterion) and encounters some new problems (e.g. sparse feature and boundary data fusion). Therefore, the work of BPRE based on network traces deserves further investigations.

6 Acknowledgments

Our work was supported by National Key Basic Research Program of China and the authors deliver grateful thanks to Associate Professor Qing Li, Professor Hongyi Yu, B.Eng. Qi Zhang, B. Eng. Xueyang Bai, M.Phil. Xiaoyong Yan and M.Phil. Hengbin Cai for selfless help and important comments.

7 References

1 Pan, F., Wu, L.F., Du, Y., *et al.*: 'Overviews on protocol reverse engineering', *Appl. Res. Comput.*, 2011, **28**, (8), pp. 2801–2806

2 Trifilo, A., Burschka, S., Biersack, E.: 'Traffic to protocol reverse engineering'. Proc. of the 2009 IEEE Symp. on Computational Intelligence in Security and Defense Applications (CISDA), Ottawa, Canada, July 2009, pp. 1–8

3 Paxson, V.: 'Bro: a system for detecting network intruders in real-time', *Comput. Netw.*, 1999, **31**, (23), pp. 2435–2463

4 Wang, H.J., Guo, C., Simon, D.R., *et al.*: 'Shield: vulnerability-driven network filters for preventing known vulnerability exploits', *ACM SIGCOMM Comput. Commun. Rev.*, 2004, **34**, (4), pp. 193–204

5 Li, X.D., Li, C.: 'A survey on methods of automatic protocol reverse engineering'. Proceedings of 2011 Seventh Int. Conf. on Computational Intelligence and Security (CIS), Sanya, China, December 2011, pp. 685–689

6 'Intemet netflow statistics – Internet2 NetFlow organization', available at http://www.internet2.edu/presentations/fall-03/20031013-NetFlow-Shalunov.pdf, accessed October 2003

7 Dagon, D., Gu, G., Lee, C.P., *et al.*: 'A taxonomy of botnet structures'. Proc. 23rd Annual Computer Security Applications Conference (ACSAC), Miami Beach, Florida, December 2007, pp. 325–339

8 Meng, F., Liu, Y., Zhang, C., *et al.*: 'Inferring protocol state machine for binary communication protocol'. Proc. of 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA), Ottawa, Ontario, Canada, September 2014, pp. 870–874

9 'How samba was written – Tridgell, A.', available at http://samba.org/ftp/tridge/misc/french_cafe.txt, accessed October 2010

10 'The Protocol Informatics Project – Marshall, A. B.', available at http://www.4tphi.net/%7eawalters/PI/PI.html, accessed March 2014

11 Cui, W.D., Kannan, J., Wang, H.J.: 'Discoverer: automatic protocol reverse engineering from network traces'. Proc. of Usenix Security Symp., Boston, MA, August 2007, pp. 199–212

12 'Security evaluation of communication protocols in common criteria using netzob – Georges, B.', available at http://www.yourcreativesolutions.nl/ICCC13/p/Networkdevices/GeorgesBossert–SecurityEvaluationofCommunicationProtocols in Common Criteria.pdf, accessed July 2014

13 Georges, B., Frédéric, G., Guillaume, H.: 'Towards automated protocol reverse engineering using semantic information'. Proc. Ninth ACM Symp. on Information, Computer and Communications Security, Kyoto, Japan, June 2014, pp. 51–62

14 Li, M., Yu, S.Z.: 'Noise-Tolerant and optimal segmentation of message formats for unknown application-layer protocols', *J. Softw.*, 2013, **24**, (3), pp. 604–617

15 Luo, J.Z., Yu, S.Z.: 'Position-based automatic reverse engineering of network protocols', *J. Netw. Comput. Appl.*, 2013, **36**, (3), pp. 1070–1077

16 Zhang, Z., Zhang, Z., Lee, P.P.C., *et al.*: 'Toward unsupervised protocol feature word extraction', *IEEE J. Sel. Areas Commun.*, 2014, **32**, (10), pp. 1894–1906

17 The Internet Engineering Task Force: 'RFC 935: Reliable link layer protocols', January 2014

18 Li, W.C., Zhou, Y., Xia, S.X.: 'A novel clustering algorithm based on hierarchical and *k*-means clustering'. Proc. 26th Chinese Control Conf. (CCC), Hunan, China, July 2007, pp. 605–609

19 International Telecommunications Union: 'Technical characteristics for an automatic identification system using TDMA in the VHF maritime mobile band, Recommendation ITU-R M.1371-4', 2010

20 International Organization for Standardization: 'Information technology-Telecommunications and information exchange between systems-High-level data link control (HDLC) procedures, ISO/IEC 13239:2002', 2007

21 The Internet Engineering Task Force: 'RFC 1088: a standard for the transmission of IP datagrams over NetBIOS networks', February 1989

22 The Internet Engineering Task Force: 'RFC 792: Internet control message protocol', September 1981