УДК 004.056

Методика лексической разметки структурированных бинарных данных сетевого трафика для задач анализа протоколов в условиях неопределенности

Гайфулина Д. А., Котенко И. В., Федорченко А. В.

Актуальность работы: *эффективность исследования сетевых протоколов для сложных информационно-коммуникационных инфраструктур, таких как киберфизические системы или "Интернет вещей", снижается по причине неопределенности структуры хранимых и передаваемых в них данных. Таким образом возникает необходимость определения спецификаций подобных нерегламентированных протоколов. Исходными данными при этом служат бинарные данные, содержащие сообщения протоколов, например, сетевой трафик в виде последовательности атомарных единиц информации (байт, бит).* ***Целью работы*** *является преодоление лексической неопределенности бинарных данных сетевого трафика на основе разметки трафика по лексемам (полям) сообщений и формирования общих лексических структур (протоколов). Предлагается подход к выявлению типовых структур протоколов формирования сообщений в бинарных данных и определения их лексических спецификаций.* ***Используемые методы:*** *в основу подхода заложены статистические методы анализа информации на естественных языках, широко применяемые при интеллектуальном анализе текстов, что является основным теоретическим вкладом данной работы.* ***Новизна:*** *лексическое распознавание условно структурированных бинарных данных на основе частотного анализа возможных последовательностей информационных единиц и их комбинаций.* ***Результат:*** *разработана методика анализа структурированных бинарных данных, апробированная на различных наборах данных сетевого трафика в условиях неопределенных спецификаций сетевых протоколов, представленных в виде упорядоченной последовательности цифровых (бинарных) данных. По результатам экспериментальной оценки точности разработанной методики, выявленные типовые последовательности (лексемы) единиц информации (байт) в среднем на 95% соответствуют содержанию полей сетевых протоколов трафика, а выявленные структуры (последовательности лексем) на 70% соответствуют спецификациям данных протоколов.* ***Практическая значимость:*** *возможность выполнения предварительной обработки структурированных бинарных данных для автоматизированного решения задач обнаружения аномалий, выполнения аудита безопасности, выявления уязвимостей протоколов, проведения тестирования приложений, а также расследования инцидентов информационной безопасности в условиях неопределенных спецификаций протоколов взаимодействия в условиях неопределенности.*

Ключевые слова: *восстановление форматов данных, структурный анализ, бинарные данные, проприетарные протоколы, анализ защищенности, неопределенные инфраструктуры.*

Введение

Существующие информационно-коммуникационные инфраструктуры отличаются сложной архитектурой, высокой гетерогенностью и большими объемами хранимой, передаваемой и обрабатываемой информации. К системам с

Библиографическая ссылка на статью:
Гайфулина Д. А., Котенко И. В., Федорченко А. В. Методика лексической разметки структурированных бинарных данных сетевого трафика для задач анализа протоколов в условиях неопределенности // Системы управления, связи и безопасности. 2019. № 4. С. 280-299. DOI: 10.24411/2410-9916-2019-10411.

Reference for citation:
Gaifulina D. A., Kotenko I. V., Fedorchenko A. V. A Technique for Lexical Markup of Structured Binary Data for Problems of Protocols Analysis in Uncertainty Conditions. *Systems of Control, Communication and Security*, 2019, no. 4, pp. 280-299. DOI: 10.24411/2410-9916-2019-10411 (in Russian).

подобной инфраструктурой относятся киберфизические системы, сети «Интернета вещей» (Internet of Things, IoT), автоматизированные системы управления технологическими процессами (АСУ ТП), системы управления транспортом и многие другие. Эффективность анализа защищенности подобных систем значительно снижается из-за высокой разнородности источников информации, несогласованности их форматов представления данных, использования нерегламентированных протоколов передачи данных и спецификаций форматов проприетарных хранилищ, а также возрастающей сложности потенциальных угроз и специально реализованных атак и вредоносного программного обеспечения [1, 2].

Поиск и устранение различных уязвимостей в среде передачи информации, аппаратном, программном и программно-аппаратном обеспечении являются актуальными задачами обеспечения безопасности, особенно в неопределенных инфраструктурах. Исходными данными анализа при этом служат бинарные («сырые») данные, содержащие сообщения различных протоколов. Под протоколом понимается набор правил, по которым осуществляется внутренний или внешний обмен данными и формируются сообщения для его осуществления между функциональными блоками (программами, файловыми системами, узлами сети и т.д.). Таким образом, исследование протоколов включает в себя задачу восстановления форматов и структур бинарных данных с целью определения алгоритмов формирования сообщений [3]. Для обозначения объекта, формат которого требуется восстановить, будет использоваться общий термин «сообщение», обозначающий сетевое сообщение или файл. С точки зрения передачи информации по каналу связи сообщение представляет собой последовательность атомарных единиц информации (байт, бит). Спецификацией сообщения является информация о типах данных, синтаксисе и семантике полей его структуры, а также о группировке полей в определенные последовательности.

Восстановленные спецификации данных применяются для тестирования программного обеспечения при статистическом и динамическом анализе бинарных последовательностей. Наличие описания структуры входных данных позволяет значительно повысить эффективность и снизить время автоматического тестирования программ для своевременного обнаружения уязвимостей, недекларированных возможностей и программно-аппаратных закладок, которые могут эксплуатироваться злоумышленниками. Спецификации форматов данных могут также использоваться в системах обнаружения вторжений для идентификации атак, в системах глубокого инспектирования пакетов (Deep Packet Inspection, DPI) для фильтрации трафика, в системах предотвращения утечек для выявления запрещенного содержимого бинарных файлов, в антивирусных решениях для обнаружения вредоносного программного обеспечения. Место и роль задачи восстановления форматов данных представлены на рис. 1.

Восстановление форматов данных при наличии кода исследуемой программы основывается на идентификации переменных программной среды (исполняемого кода, окружения и др.). Для получения информации о процессе разбора сообщения применяется метод формирования трассы в процессе программной обработки сообщения внутри эмулятора. Такой метод реализует ди-

№4. 2019
ISSN 2410-9916

намический анализ программ и позволяет восстановить большое количество семантической информации благодаря наблюдению за процессом обработки сообщения. Однако при отсутствии доступа к коду программы, например, в случае выполнения кода в виде удаленного сервиса, восстановление форматов данных выполняется преимущественно в ручном режиме, что требует высокой квалификации аналитика и значительных временных затрат. По этой причине автоматизация задачи восстановления форматов данных в условиях неопределенности является актуальным направлением исследований.

В данном случае неопределенность выражается как в отсутствии объекта обработки сообщения (кода программы), так и невозможности тривиального определения спецификации его данных (полей бинарных форматов). В свою очередь, поля спецификаций данных могут быть неопределенными: (1) лексически – при отсутствии разделения данных по структурным единицам – лексемам; (2) синтаксически – при отсутствии данных о правилах присвоения значений и функциональном взаимодействии лексем; (3) семантически – при отсутствии данных о смысловых нагрузках лексем.

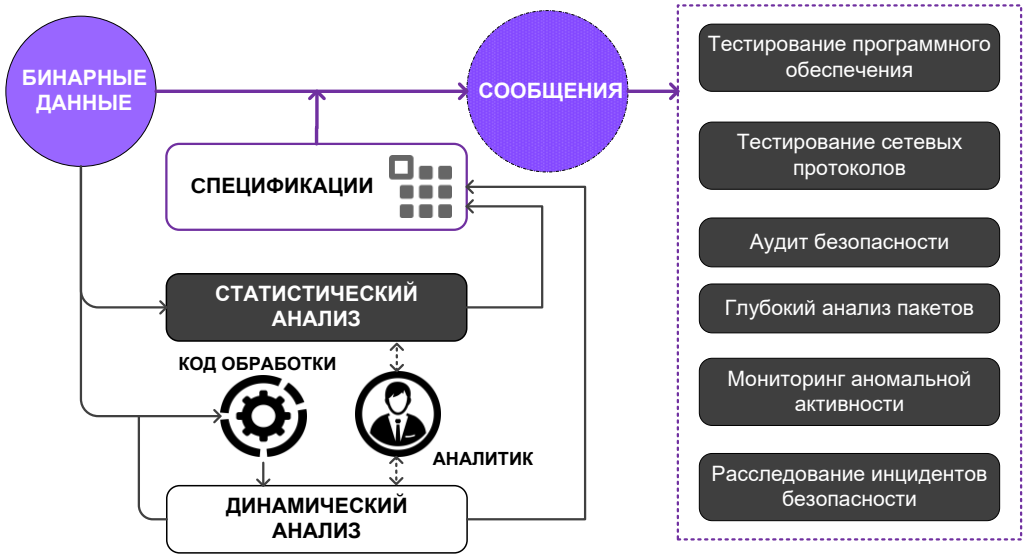

Рис. 1. Место и роль анализа форматов бинарных данных

Текущее исследование направлено на преодоление лексической неопределенности путем разметки трафика по лексемам (полям) сообщений и формирования общих лексических структур (протоколов). В основу методики заложены статистические подходы анализа информации на естественных языках, широко применяемые при интеллектуальном анализе текстов, что является основным теоретическим вкладом данной работы. Научной новизной подхода является лексическое распознавание условно структурированных бинарных данных на основе частотного анализа возможных последовательностей информационных единиц и их комбинаций. Практическая значимость исследований заключается в преодолении лексической неопределенности данных с целью выполнения различных задач обеспечения безопасности, таких как выявление сетевых аномалий, идентификация вредоносного контента, оценка защищенно-

DOI: 10.24411/2410-9916-2019-10411
URL: https://sccs.intelgr.com/archive/2019-04/11-Gaifulina.pdf

сти, проведение аудита безопасности и расследование инцидентов в неопределенных и гетерогенных средах обработки данных (например, в киберфизических системах и сетях «Интернета вещей»).

Таким образом, целью текущего исследования является *классификация и определение спецификаций структурированных бинарных («сырых») данных сетевого трафика в условиях неопределённости протоколов передачи информации для повышения эффективности последующего выделения сетевых потоков, представляющих угрозу.*

В статье развиваются ранее полученные результаты, основные положения которых описаны в [4]. Данная статья демонстрирует применение разрабатываемого авторами подхода для более широкой области анализа бинарных данных, в частности, вносятся новые положения и обозначения используемых единиц и переменных, что позволяет перейти к определению форматов данных для различных протоколов. Приводится развернутое описание предлагаемого подхода и проводимых экспериментов по анализу структурированных бинарных данных. В качестве исходных данных для эксперимента используются новые записи сетевого трафика, специфицирующие использование проприетарных протоколов киберфизических систем, а также представляются новые способы визуализации результатов эксперимента.

Работа организована следующим образом. Вначале проводится анализ релевантных работ, посвященных определению спецификаций бинарных форматов данных (в том числе, сетевых протоколов) и поиску ключевых слов в текстовых данных. Затем описывается методика определения формального приближения структур бинарных данных в условиях неопределенности и требования к исходным данным. В следующем разделе рассматриваются результаты экспериментов по анализу структурированных бинарных данных сетевого трафика, преимущества предлагаемого подхода, причины некорректного выполнения разработанной методики, а также возможные способы их преодоления.

Анализ работ в исследуемой предметной области

Существует несколько различных подходов к восстановлению формата сообщений, среди которых можно выделить две основные группы:

1) статистический анализ данных (например, сетевого трафика или бинарных файлов) с выделением повторяющихся паттернов;
2) динамический анализ кода одного из участников обмена сообщениями.

Процесс восстановления форматов данных с одной стороны является самостоятельной и важной задачей обратной инженерии, с другой стороны, в рамках анализа сетевого трафика, он является частью задачи анализа сетевых протоколов. Анализ сетевого трафика главным образом основывается на разборе заголовков протоколов в пакетах. В зависимости от количества протоколов, а также глубины их стека и уровня в модели OSI (Open System Interconnection), выделяют три основных вида анализа трафика:

1) поверхностный (Shallow Packet Inspection, SPI);
2) средний (Medium Packet Inspection, MPI);

3) глубокий (Deep Packet Inspection, DPI).

Очевидно, что наиболее сложным и результативным видом анализа является DPI [5], реализующий накопление статистических данных о сетевых потоках, а также проверку и фильтрацию сетевых пакетов. Описанный вид анализа применяется в одноименном классе систем (DPI), реализующих правило-ориентированные и интеллектуальные методы анализа трафика. В системах анализа защищенности среды передачи данных выполняется разбор сетевых пакетов с использованием технологии DPI для оценки потенциальной опасности их содержимого. Для применения указанной технологии также необходимо описание структуры сетевых пакетов.

Использование статистических подходов к восстановлению форматов сообщений заключается в извлечении полей сообщения в виде ключевых слов и дальнейшее проведение их синтаксического анализа. Примерами инструментов, реализующих указанные подходы, являются Biprominer [6], ProDecoder [7], AutoReEngine [8], ReverX [9]. Статистические походы анализа данных осуществляются в три этапа:

1) идентификация статистически-значимых паттернов (шаблонов) данных в качестве ключевых слов;
2) определение сообщений с помощью отличительных ключевых слов;
3) определение порядка и взаимосвязи ключевых слов в данных для поиска вероятных последовательностей сообщений.

Выделяют статистические и структурные методы извлечения ключевых слов сообщений. Статистические методы извлечения ключевых слов учитывают относительные частоты встречаемости языковых единиц и их комбинаций. Одним из классических методов в данном классе является расчет для каждого слова меры TF-IDF (Term Frequency-Inverse Document Frequency) [10], которая отражает его важность в конкретном тексте среди коллекции документов.

Структурные методы анализа ключевых слов представлены граф-ориентированными подходами, например, Rake [11] и DegExt [12]. В данных алгоритмах строится граф, вершинами которого выступают отдельные слова. Значимость слов характеризуется набором показателей, таких как частота наблюдения слова, связность с другими словами (степень) и их отношение. Стоит отметить, что статистические инструменты анализа не требуют доступа к коду программного обеспечения, реализующего протокол взаимодействия, и поэтому являются подходящей отправной точкой для администратора сети, который исследует подозрительный сетевой трафик. При этом статистический анализ позволяет преодолеть лексическую и синтаксическую неопределенность данных, в то время как основным недостатком остается сложность определения семантики полей сообщений.

Методики динамического анализа данных, или фаззинга, контролируют выполнение кода программного обеспечения, реализующего протокол взаимодействия, и осуществляют выявление используемых спецификаций бинарных данных. Техника фаззинга позволяет отслеживать инструкции, которые вызываются при выполнении кода (программного или машинного) на специально сформированных либо случайных исходных данных [13]. Наблюдая за выпол-

нением конкретных инструкций по обмену сообщениями и использованием памяти участников обмена, инструменты анализа выводят формат сообщений и, в некоторых случаях, их семантику.

В ряде работ описаны методики динамического анализа бинарных файлов. В [14] представлен инструмент Dispatcher, использующих трехфазную методику анализа:

1) выполнение программного кода с искажением передаваемых на вход данных («заражением»);
2) идентификация границ полей фиксированной и переменной длины;
3) выделение ключевых слов.

В инструменте ReFormat [15] применяется следующий двухфазный подход:

1) идентификация наборов команд шифрования;
2) деконструирование исследуемого сообщения в источники динамического «заражения».

В работе [16] предлагается подход, при котором сообщения выстраиваются в одну из двух иерархических категорий: по восходящим или по нисходящим грамматикам. Нисходящие грамматики сообщений анализируют сообщения на основе информации предыдущих значений полей (заголовков сообщений), в то время как восходящие грамматики анализируют сообщения от полезной нагрузки, не полагаясь на заголовок. Описанный инструмент наблюдает за направлением последовательности инструкций программы, отслеживает чтение или запись данных в буфере памяти и восстанавливает структуру сообщения в виде дерева полей данных. Основным ограничением динамических подходов к анализу данных становится зависимость результата от тестируемой реализации программного обеспечения. Если реализация протокола игнорирует определенные поля сообщения, тогда эти поля не будут выведены [17]. Также недостатком динамических подходов к анализу бинарных файлов является восстановление только той части структуры сообщения, которая была обработана при выполнении на фиксированных входных данных (конкретном сетевом сообщении или файле).

Таким образом, техника фаззинга реализует анализ бинарных данных методом «серого ящика», когда у исследователя имеется как минимум машинный код обработки информации. В свою очередь, предлагаемый в настоящей работе подход реализует анализ бинарных данных методом «черного ящика», то есть методика не принимает дополнительных сведений, кроме самой последовательности «сырых» данных. Применяемый в работе подход структурного анализа бинарных данных основан на методах и алгоритмах извлечения ключевых слов из текстов, написанных на естественных языках.

Для анализа структуры бинарных данных также применяются подходы, использующие методы машинного обучения и интеллектуального анализа данных для извлечения ключевых слов. Подходы на основе методов для создания обучающей выборки и построения модели-классификатора, такие как байесовские методы [18], метод опорных векторов [19], деревья решений [20] и нейронные сети [21], как правило, требуют обучающей выборки сообщений с

размеченными ключевыми словами. Преимуществами чисто статистического подхода являются универсальность алгоритмов извлечения ключевых слов и отсутствие необходимости в трудоемких и ресурсозатратных процедурах построения баз знаний.

Методика лексической разметки

Для разметки бинарных данных с неизвестными спецификациями форматов (протоколов) сообщений необходимо преодоление лексических неопределенностей разных уровней, заключенных в отсутствии сведений о: (1) границах сообщений в исходных данных; (2) структурах сообщений (заголовков и полезной нагрузке); (3) структурах заголовков сообщений в виде упорядоченного набора полей. Таким образом, результатом лексического восстановления спецификаций бинарных форматов данных является формальное приближение заголовков сообщений, а также их полей, к исходным структурам соответствующих протоколов. Исходными данными для выполнения предлагаемой методики анализа является структурированная последовательность B единиц информации (в описываемом случае, мы ограничиваемся байтами данных), отражающая ряд непересекающихся сообщений.

Первый этап методики заключается в разбиении последовательности B размерностью N байт на всевозможные подпоследовательности db (слова, или лексемы) с размерностями k байт для создания словаря DW:

$$B = \{db_1, db_2, ..., db_e\},\ e = \left\lfloor \frac{N}{2} \right\rfloor + \left\lfloor \frac{N}{3} \right\rfloor + ... + \left\lfloor \frac{N}{k_{\max}} \right\rfloor, \quad (1)$$

$$db = \{b_1, ..., b_k\}, k \in [2, k_{\max}], k_{\max} \le N, \quad (2)$$

где k_{max} – максимальная длина лексемы, а e – размер словаря.

Графическая иллюстрация первого этапа методики представлена на рис. 2. На данном этапе выделяется k_{max} параллельных потоков обработки исходных данных. Каждый поток принимает на вход свое уникальное значение $k \in [2, k_{\max}]$, осуществляет разбиение исходной последовательности байт на отрезки длиной k и определяет количество повторяющихся отрезков. Далее окно поиска смещается с длиной шага $\Delta \in [0, k-1]$ от первого байта в последовательности. Таким образом, каждый поток совершает k итераций поиска повторившихся хотя бы раз последовательностей байт в исходных данных.

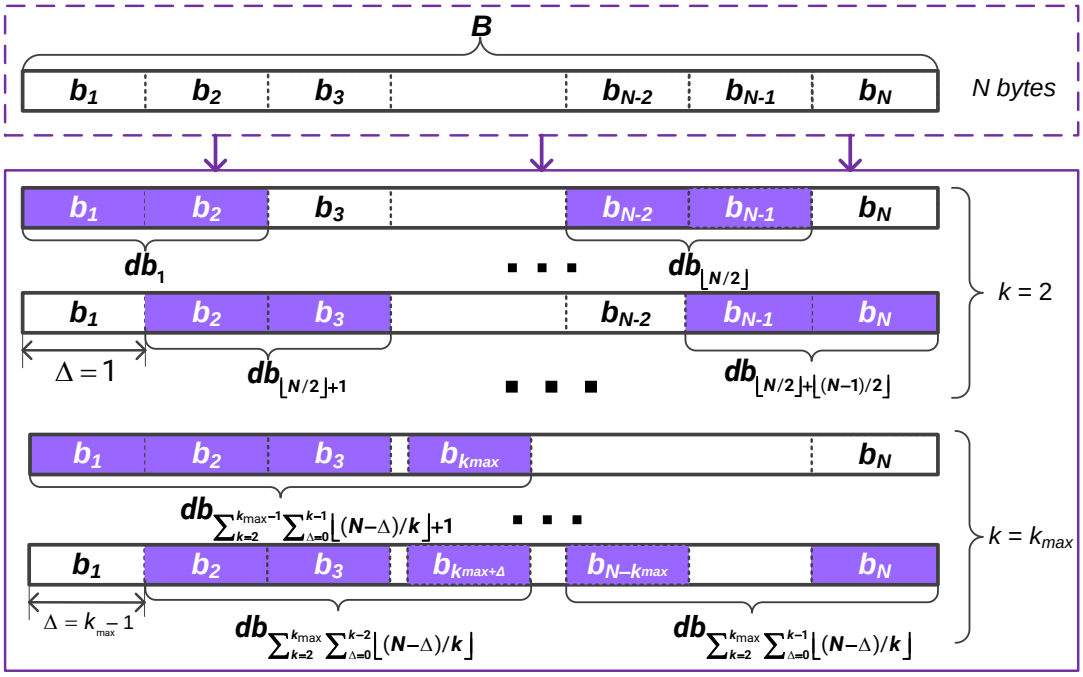

Рис. 2. Схема выделения подпоследовательностей бинарных данных

В сформированном словаре каждое слово *w* описывается уникальным порядковым идентификатором *iw*, значением *vw*, длиной *k* и частотой наблюдения *cw*:

$$w = \langle iw, vw, k, cw \rangle, w \in DW, k \in [2, N]. \quad (3)$$

Второй этап методики заключается в фильтрации полученного словаря *DW*. Смысл данного этапа заключается в выделении значимых лексем. К таким лексемам относятся подпоследовательности, наблюдаемые не менее двух раз, при этом их границы не должны пересекаться в последовательности *B*. Под пересечением в данном случае понимается принадлежность очередного неразмеченного байта *s* к двум и более словам (w_x и w_y). В упорядоченную последовательность слов *LI* добавляется очередной идентификатор слова (iw_x или iw_y). с наибольшей частотой наблюдения (cw_x или cw_y), а остальные слова исключаются. Каждая позиция байт исходных данных, которая в результате анализа не причислена ни к одной извлеченной типовой последовательности, соответствует нулевому идентификатору. Данный идентификатор обозначает лексическую неопределенность участка бинарных данных. Таким образом, последовательность индексов значимых слов *iw'* списка *LI* определяется по следующему принципу:

$$iw' = \begin{cases} iw_x, s \in w_x, w_y, k_x cw_x > k_y cw_y; \\ iw_y, s \in w_x, w_y, k_x cw_y < k_y cw_y; \\ 0, s \notin w_x, w_y. \end{cases} \quad (4)$$

Третий этап методики заключается в определении цепей слов бинарных данных. Каждая цепь состоит из последовательностей лексем, для каждой из которых указан размер и смещение относительно начала сообщения. Из списка

LI извлекаются возможные последовательности порядковых идентификаторов слов по следующим условиям: последовательность не должна начинаться с нулевого идентификатора и должна содержать не менее двух слов. Далее определяются идентичные друг другу последовательности, схожие по включаемым в них слова и их порядку в цепи. Длина искомой последовательности слов (количество включаемых идентификаторов) увеличивается на единицу до некоторого значения, при котором в исходных данных не будет найдено ни одной идентичной пары последовательностей. Каждая определенная цепь лексем *c* в формируемом словаре *DC* описывается порядковым идентификатором *ic*, значением *vc*, длиной *z*, а также частотой наблюдения *cc*:

$$c = \langle ic, vc, z, cc \rangle, c \in DC, z \in [2, |DW|]. \quad (5)$$

Четвертый этап методики заключается в упорядочивании цепей слов без пересечения на исходных данных. В последовательность *LS* для каждой позиции последовательности байт добавляется порядковый идентификатор цепи лексем с наибольшей частотой наблюдения. При этом нулевому идентификатору аналогично соответствуют участки данных, в которых цепи слов не определены. Структура бинарных данных *LS* по упорядоченным последовательностям лексем *ic'* определяется по следующему принципу:

$$ic' = \begin{cases} ic_x, s \in c_x, c_y, z_x cc_x > z_y cc_y; \\ ic_y, s \in c_x, c_y, z_x cc_x < z_y cc_y; \\ 0, s \notin c_x, c_y. \end{cases} \quad (6)$$

Пятым этапом методики является восстановление спецификации (структуры) сообщений в бинарных данных. Заголовки сообщений идентифицируются как наиболее часто наблюдаемые цепи слов в последовательности *LS*, полученной на *четвертом этапе*. Определяются параметры «слов» в каждой структуре поля сообщения *field',* такие как идентификатор *iw*, длина *k* и частота наблюдения среди сообщений *ct*:

$$field' = \langle iw, k, ct \rangle. \quad (7)$$

Далее определяются границы полей в структуре сообщения путем подбора и анализа разметки с использованием словаря лексем. Таким образом, структура каждого сообщения состоит из множества полей (принадлежащих конкретной выделенной структуре спецификации). В свою очередь, каждое поле сообщения *field* описывается идентификатором поля *if* и лексическим P^{lex}, синтаксическим P^{syn} и семантическим P^{sem} типом информации:

$$field = \langle if, P^{lex}, P^{syn}, P^{sem} \rangle. \quad (8)$$

Лексический тип определяет размер лексемы (поля) сообщения в бинарных данных. Множество лексических типов сообщений сформировано по размерности машинного слова классической архитектуры Intel8086 для целых значений. Данное множество включает следующие типы полей: 8-разрядное (*BYTE*), 16-разрядное (*WORD*), 32-разрядное (*DWORD*), 64-разрядное

(QWORD), а также массив *n*-байт (*nBYTES*). Семантический тип поля сообщения отражает его функциональное значение. Определение данного типа полей является наиболее трудновыполнимой и актуальной задачей, поскольку позволит автоматизировано оценивать эффективность применения протоколов, а также выявлять в них разнообразные ошибки и уязвимости. Определение семантического типа лексем планируется реализовать за счет логического вывода методами дискрипционной логики.

Данная работа направлена исключительно на определение лексической неопределенности бинарных данных. Результатом выполнения предлагаемой методики является последовательность бинарных данных, размеченная по определенным полям (лексемам) протоколов формирования сообщений. При анализе разметки сообщений может возникнуть ситуация, что размер некоторого поля варьируется в нескольких значениях. В таком случае появляются варианты структуры сообщения $bstr_x$ и $bstr_y$, для которых поле ($field_x$ или $field_y$) с одинаковым смещением от начала заголовка x имеет разный лексический тип (P_x^{lex} и P_y^{lex}):

$$P_x^{lex} \neq P_y^{lex}, x = y \Rightarrow \begin{cases} P_x^{lex} \in field_x, field_x \in bstr_x; \\ P_y^{lex} \in field_y, field_y \in bstr_y. \end{cases} \quad (9)$$

Анализируя возможные лексические типы полей сообщений, получается множество вариантов структур бинарных данных $BSTR = \{bstr\}$.

Описанный подход к анализу бинарных данных ориентирован на достижение максимального покрытия при сохранении наибольшей точности разметки.

Схема описанной методики анализа бинарных данных предоставлена на рис. 3.

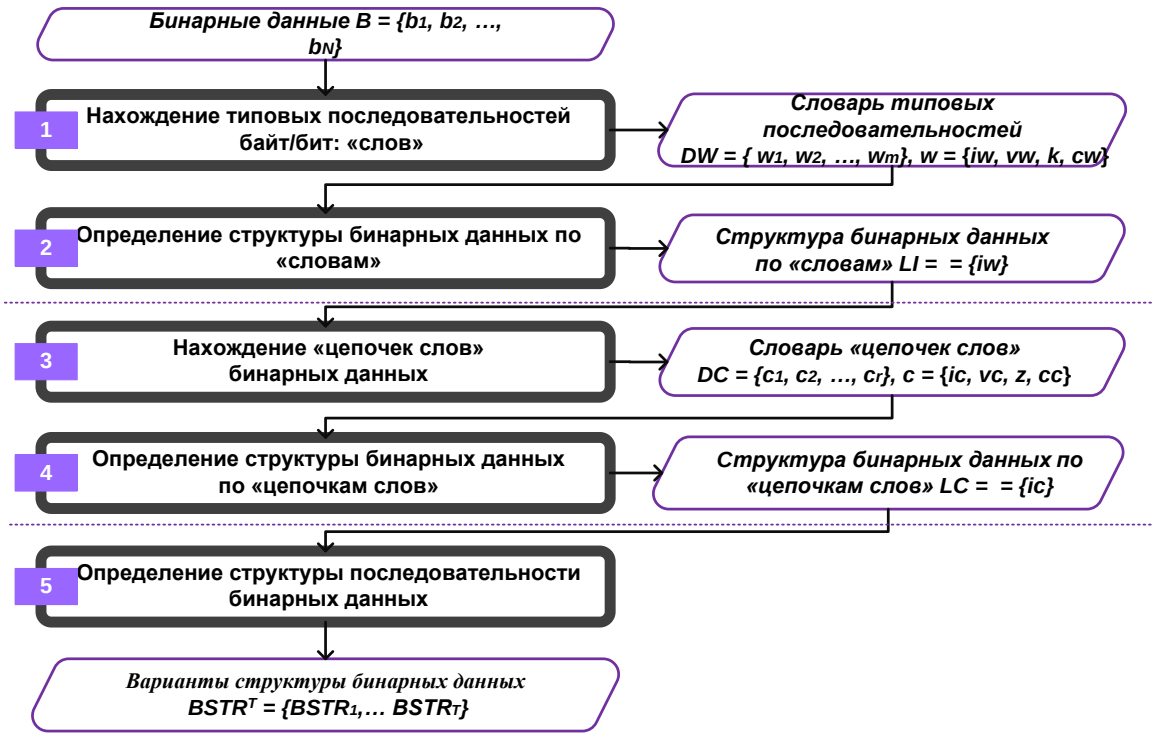

Рис. 3. Схема методики анализа бинарных данных

Для корректного выполнения разработанной методики определен ряд требований к исходным данным по следующим свойствам:

- структурированность - упорядоченность спецификаций сообщений и полей сообщений в исходных данных, то есть отсутствие их фрагментированности;
- полнота исходных бинарных данных для выполнения описанных этапов по извлечению слов и определению типовой структуры каждого из содержащихся в бинарных данных протокола;
- однотипность (целостность) структуры спецификаций, отсутствие (минимизация) шифрованных, хэшированных и рандомизированных данных, имеющих достаточно высокую уникальность.

Эксперименты

Предлагаемая методика анализа была реализована на языке программирования Python и апробирована на различных наборах бинарных данных. В качестве исходных данных для анализа использовались последовательности байт сетевого трафика в условиях неопределенных спецификаций сетевых протоколов. Протоколы являются наборами правил формирования соответствующих сетевых сообщений и очередности их отправки для выполнения конкретных задач по обеспечению работоспособности сети. Сетевые протоколы модели OSI организуются в стек таким образом, что каждый очередной протокол является полезной нагрузкой предыдущего. Сообщения конкретных протоколов, а также их стеки, формируют пакеты, которые отправляются через сетевые интерфейсы и составляют сетевой трафик. Сетевые пакеты могут иметь различную структуру, описываемую в рамках протокола, по которому происходит коммуникационная сессия в рамках поддерживаемого сетевого соединения. Однако отсутствие сведений о спецификации протокола более низкого уровня подразумевает отсутствие информации о спецификации последующих протоколов в его стеке. В условиях неопределенных спецификаций сетевых протоколов запись трафика сети за фиксированный временной период является конечным подмножеством передаваемой информации, представленной в виде упорядоченной последовательности цифровых (бинарных) данных в виде пакетов.

Результаты проведенных экспериментов по определению неопределенных спецификаций сетевых протоколов представлены в таблице 1.

В данной таблице описываются реальные характеристики анализируемого трафика, характеристики, полученные в результате выполнения разработанной методики, и сравнительная оценка (соотношения) полученных и фактических характеристик данных.

Исследуемый набор данных был сгенерирован для сценария автоматизации процессов управления с использованием оборудования MODBUS/TCP с целью исследования применения методов машинного обучения для кибербезопасности в автоматизированных системах управления [22].

Анализ результатов проведенных экспериментов показал, что методика позволяет достигнуть более высокой эффективности определения слов, совпа-

дающих с реальными полями сетевых протоколов (*DVF*) при больших количествах наблюдаемых протоколов в стеке и их значений (снимки 5-6). Наличие большего объема данных полей протоколов в трафике (*UFV*, *DD*) позволяет получить более высокую оценку точности разметки сетевого трафика по пакетам (*DP*) и полям протоколов (*DPb*). Оценка точности разметки трафика по протоколам (*DPb*), в свою очередь, демонстрирует высокую эффективность методики при наличии большего числа пакетов (*P*) в совокупности с относительно небольшим количеством используемых сетевых протоколов в трафике (снимок 1). Это обосновывается достаточным наличием дублируемых данных сетевого трафика, необходимым для анализа его структуры.

Таблица 1 – Результаты обработки сетевого трафика с помощью предлагаемой методики анализа

№ снимка	Размер, МБ	Фактическая структура трафика			Структура нерегламентированных протоколов					Оценка результатов			
		Уникальные значения полей, N	*Пакеты, N*	*Повторяющиеся данные, байты*	*Слова, N*	*Найдено полей протоколов, N*	*Определено пакетов, N*	*Данные полей, байты*	*Данные протоколов, байты*	*DVF / UVF*	*P / DP*	*DFb / DD*	*DPb / DD*
		UVF	***P***	***DD***	***W***	***DVF***	***DP***	***DFb***	***DPb***				
1	2,78	124013	35430	2,55	15961	79368	34721	2,45	1,938	0,64	0,98	0,96	0,76
2	5,74	207942	72150	5,3	36858	149718	63492	4,99	4,346	0,72	0,88	0,94	0,82
3	8,39	298247	105508	7,56	52079	196843	100232	7	4,914	0,66	0,95	0,93	0,65
4	11,3	320991	142194	10,15	71577	189 384	130818	9,5	6,902	0,59	0,92	0,94	0,68
5	2,78	116549	35477	2,1	15879	88577	33703	1,9	1,512	0,76	0,95	0,93	0,72
6	5,56	211879	70863	4,87	31840	131364	68737	4,6	3,7499	0,62	0,97	0,95	0,77
7	2,78	118950	35430	2,28	15838	131364	30115	2,2	1,5732	0,74	0,85	0,97	0,69
8	5,44	229128	69440	4,76	30935	155 807	63884	4,57	3,808	0,68	0,92	0,96	0,8
Среднее значение										**0,67**	**0,93**	**0,95**	**0,74**

На рис. 4 представлен пример визуализации результатов лексической разметки сетевого трафика после определения начала заголовков сообщений (пакетов). Каждая ячейка является байтом данных, последовательность байтов следует слева на право, а последовательность сообщений (пакетов Ethernet) – сверху вниз. Номер ячейки определяет принадлежность текущей информационной единицы (байта) к определенному полю (лексеме) конкретного протокола. Дублирующиеся в пакетах трафика байты полей протоколов помечены одинаковыми номерами и выделены одинаковыми цветами. В рамках данного примера, каждое поле всех протоколов имеет уникальный цвет и порядковый номер.

Рис. 4. Пример лексической разметки сетевого трафика по сообщениям и выделенным «словам»

На данном изображении можно отметить, что лексемы расположены в сообщении разнородно, не формируя единообразную структуру. Далее проводится сравнение сообщений по последовательности лексем с использованием множества слов. Результатом данного анализа являются возможные варианты разметки структуры сообщений в сетевом трафике.

На рис. 5 представлен пример визуального представления спецификаций сообщений. Выделенные лексемы (слова) по отношению к исходным данным, изображенным на рис. 3, имеют аналогичный принцип отображения. Наименование ячейки определяет принадлежность байта к определенному полю в восстановленной структуре сообщения.

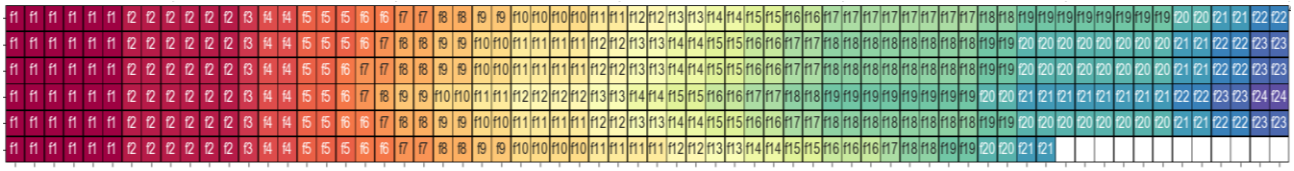

Рис. 5.Пример различных вариантов разметки сообщений
сетевого трафика по полям протоколов (последовательностям лексем)

Таким образом, в результате анализа было восстановлено пять возможных спецификаций сообщений в сетевом трафике. Для каждой спецификации проведено сравнение с фактической структурой сообщения по полям сетевых протоколов и определено их совпадение в интервале от 59% до 76%. На рис. 6 отображен пример сравнения полей сетевого сообщения по ее фактической структуре и восстановленной структуре с использованием предлагаемой методики.

В фактической структуре сообщения каждая строка отображает поле сетевого протокола и ее лексический тип. В восстановленной спецификации аналогично отображены лексические типы полей сообщений.

Сопоставление данных полей отображено в четырех видах:

- сплошная двойная стрелка – полное совпадение полей (например, для MAC-адресов, IP-адресов);
- стрелка слева-направо – слияние фактических полей сообщения (для типа обслуживания и длины пакета);
- стрелка справа-налево – дробление фактического поля сообщения (для порядкового номера);
- пунктирная двойная стрелка – отсутствие совпадения (для Ethernet типа, версии IP протокола и длины пакета).

Рис. 6. Пример оценки точности восстановления структуры сообщения

Можно отметить, что полное совпадение полей является наиболее удачным результатом и достигается при наличии достаточного числа дублируемых данных в полях сетевого трафика. Слияние нескольких полей фактической структуры сообщения в одно поле спецификации происходит по причине следования друг за другом полей с малой длиной. Дробление поля фактической структуры сообщения проявляется при высокой частоте наблюдения его элементов в других полях, отчего части значения поля воспринимаются как отдельные ключевые слова. Слияние нескольких полей является более критичным для конечного результата, так как приводят к потере значимости полей фактической структуры как отдельной и самостоятельной единицы для последующего анализа. Дробление же может являться одним из способов детализации содержания полезной нагрузки сообщения, что позволит выделить более подробную структуру бинарных данных. Отсутствие совпадения в восстановленных полях является наиболее критичной ошибкой, которая может привести к пересечению полей разных протоколов (в примере это показано для протоколов Ethernet и IP).

На рис. 7 представлен пример визуального представления исходных данных сетевого трафика в виде набора сообщений (пакетов), размеченных по структуре, наиболее совпадающей с фактической спецификацией. Выделенные лексемы типовых структур сообщений в сетевом трафике имеют аналогичные принципы отображения, как и на рис. 4-5.

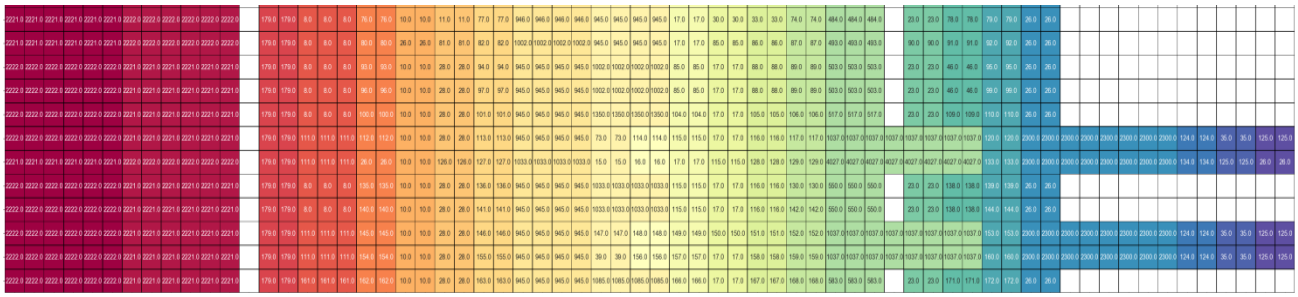

Рис. 7. Пример лексической разметки структуры
сообщений сетевого трафика

Заключение

В результате проведенных исследований разработана методика анализа структурированных бинарных данных. Предлагаемый подход основан на применении методов структурного анализа неразмеченных («сырых») данных. Определены требования к исходным данным, ограничивающие применение разработанной методики.

Подход к анализу бинарных данных был апробирован на различных наборах данных сетевого трафика в условиях неопределенных спецификаций сетевых протоколов, представленных в виде упорядоченной последовательности цифровых (бинарных) данных. По результатам экспериментальной оценки точности разработанной методики выявленные типовые последовательности (лексемы) единиц информации (байт) в среднем на 95% соответствуют содержанию полей сетевых протоколов трафика, а выявленные структуры (последовательности лексем) на 70% соответствуют спецификациям данных протоколов. Данные оценки на новых записях сетевого трафика соответствуют полученным результатам ранее проведенных экспериментов [4], что подтверждает применимость разрабатываемой методики. Также продемонстрировано улучшение прежних показателей эффективности разметки трафика.

Низкие показатели оценки точности определения протоколов сетевого трафика в ряде случаев являются следствием недостаточного количества дублируемых данных сетевого трафика, необходимым для анализа его структуры.

Практическая значимость работы заключается в возможности выполнения предварительной обработки структурированных бинарных данных для автоматизированного решения задач обнаружения аномалий, выполнения аудита безопасности, выявления уязвимостей протоколов, проведения тестирования приложений, а также расследования инцидентов информационной безопасности в условиях неопределенных спецификаций протоколов взаимодействия в условиях неопределенности.

Дальнейшим направлением исследования является развитие предложенного подхода для определения синтаксических и семантических типов полей сообщений и их спецификаций. В работе по данному направлению планируется также адаптировать методику разметки бинарных данных к параллельной обработке больших исходных данных.

Работа выполнена при частичной финансовой поддержке РФФИ (проекты 18-07-01488 и 18-29-22034) и бюджетной темы 0073-2019-0002.

Литература

1. Kotenko I., Stepashkin M. Network Security Evaluation based on Simulation of Malefactor's Behavior // SECRYPT 2006. International Conference on Security and Cryptography. Proceedings. Portugal. 7–10 August 2006. P. 339-344.

2. Komashinskiy D., Kotenko I. Malware Detection by Data Mining Techniques Based on Positionally Dependent Features // Proceedings of the 18th Euromicro Conference on Parallel, Distributed and Network-Based Processing, PDP 2010. IEEE, 2010. P. 617-623.

3. Гетьман А. И., Маркин Ю. В., Падарян В. А., Щетинин Е. И. Восстановление формата данных // Труды Института системного программирования РАН. 2010. Т. 19. С. 195-214.

4. Гайфулина Д.А. Методика анализа трафика среды передачи данных в киберфизических системах для выявления сетевых аномалий // Альманах научных работ молодых ученых XLVIII научной и учебно-методической конференции Университета ИТМО. 2019. Т. 2. С. 14-19.

5. Bremler-Barr A., Harchol Y., Hay D., Koral Y. Deep packet inspection as a service // Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies. ACM, 2014. P. 271-282.

6. Wang Y., Li X., Meng J., Zhao Y., Zhang Z., Guo L. Biprominer: Automatic mining of binary protocol features // Proceedings of the 12th International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE, 2011. P. 179-184.

7. Wang Y., Yun X., Shafiq M. Z., Wang L., Liu A. X., Zhang Z., Yao D., Zhang Y., Guo L. A semantics aware approach to automated reverse engineering unknown protocols // Proceedings of the 20th IEEE International Conference on Network Protocols. IEEE, 2012. P. 1-10.

8. Luo J. Z., Yu S. Z. Position-based automatic reverse engineering of network protocols // Journal of Network and Computer Applications. 2013. Vol. 36. № 3. P. 1070-1077.

9. Antunes J., Neves N., Verissimo P. Reverse engineering of protocols from network traces // Proceedings of the 18th Working Conference on Reverse Engineering. IEEE, 2011. P. 169-178.

10. Sparck J. K. A statistical interpretation of term specificity and its application in retrieval // Journal of documentation. 1972. Vol. 28№. 1. P. 11-21.

11. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents // Text mining: applications and theory. 2010. Vol. 1. P. 1-20.

12. Litvak M., Last M., Kandel A. Degext: a language-independent keyphrase extractor // Journal of Ambient Intelligence and Humanized Computing. 2013. Vol. 4. № 3. P. 377-387.

13. Caballero J., Song D. Automatic protocol reverse-engineering: Message format extraction and field semantics inference // Computer Networks. 2013. Vol. 57. № 2. P. 451-474.

14. Caballero J., Poosankam P., Kreibich C., Song D. Dispatcher: Enabling active botnet infiltration using automatic protocol reverse-engineering // Proceedings of the 16th ACM conference on Computer and communications security. ACM, 2009. P. 621-634.

15. Wang Z., Jiang X., Cui W., Wang X., Grace M. ReFormat: Automatic reverse engineering of encrypted messages // European Symposium on Research in Computer Security. Springer, Berlin, Heidelberg, 2009. P. 200-215.

16. Lin Z., Zhang X., Xu D. Reverse engineering input syntactic structure from program execution and its applications // IEEE Transactions on Software Engineering. 2009. Vol. 36. №. 5. P. 688-703.

17. Narayan J., Shukla S. K., Clancy T. C. A survey of automatic protocol reverse engineering tools // ACM Computing Surveys (CSUR). 2016. Vol. 48. № 3. P. 40:1-40:26.

18. Wang C., Li S. CoRankBayes: Bayesian learning to rank under the co-training framework and its application in keyphrase extraction // Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011. P. 2241-2244.

19. Isa D., Lee L. H., Kallimani V. P., Rajkumar R. Text document preprocessing with the Bayes formula for classification using the support vector machine // IEEE Transactions on Knowledge and Data engineering. 2008. Vol. 20. № 9. P. 1264-1272.

20. Jiang X., Hu Y., Li H. A ranking approach to keyphrase extraction // Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009. P. 756-757.

21. Lopez P., Romary L. HUMB: Automatic key term extraction from scientific articles in GROBID // Proceedings of the 5th international workshop on semantic evaluation. – Association for Computational Linguistics, 2010. P. 248-251.

22. Frazão. I., Abreu P., Cruz T. Araujo A., Simoes P. Cyber-security Modbus ICS dataset // IEEE Dataport [Электронный ресурс], 2019. DOI: 10.21227/pjff-1a03.

References

1. Kotenko I., Stepashkin M. Network Security Evaluation based on Simulation of Malefactor's Behavior. *SECRYPT 2006. International Conference on Security and Cryptography.* Proceedings. Portugal. 7–10 August 2006, pp. 339-344.

2. Komashinskiy D., Kotenko I. Malware Detection by Data Mining Techniques Based on Positionally Dependent Features. *Proceedings of the 18th Euromicro Conference on Parallel, Distributed and Network-Based Processing, PDP 2010.* IEEE, 2010. pp.617-623.

3. Getman A. I., Markin Yu. V., Padaryan V. A., Schetinin E. I. Data format recovery. *Proceedings of the Institute for System Programming of the Russian Academy of Sciences*. 2010, vol. 19, pp. 195-214 (in Russia).

4. Gaifulina D. A. The technique of traffic analysis of the data transmission medium in cyberphysical systems to detect network anomalies. *Almanac of scientific works of young scientists of the XLVIII scientific and educational conference of ITMO University*, 2019, vol. 2, pp. 14-19 (in Russia).

5. Bremler-Barr A., Harchol Y., Hay D., Koral Y. Deep packet inspection as a service. *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*. ACM, 2014, pp. 271-282.

6. Wang Y., Li X., Meng J., Zhao Y., Zhang Z., Guo L. Biprominer: Automatic mining of binary protocol features. *Proceeding of the 12th International Conference on Parallel and Distributed Computing, Applications and Technologies*. IEEE, 2011, pp. 179-184.

7. Wang Y., Yun X., Shafiq M. Z., Wang L., Liu A. X., Zhang Z., Yao D., Zhang Y., Guo L. A semantics aware approach to automated reverse engineering unknown protocols. *Proceedings of the 20th IEEE International Conference on Network Protocols*. IEEE, 2012, pp. 1-10.

8. Luo J. Z., Yu S. Z. Position-based automatic reverse engineering of network protocols. *Journal of Network and Computer Applications*. 2013, vol. 36, no. 3. pp. 1070-1077.

9. Antunes J., Neves N., Verissimo P. Reverse engineering of protocols from network traces. *Proceedings of the 18th Working Conference on Reverse Engineering*. IEEE, 2011, pp. 169-178.

10. Sparck J. K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*. 1972, vol. 28, no. 1, pp. 11-21.

11. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents. *Text mining: applications and theory*. 2010, vol. 1, pp. 1-20.

12. Litvak M., Last M., Kandel A. Degext: a language-independent keyphrase extractor. *Journal of Ambient Intelligence and Humanized Computing*, 2013, vol. 4. no. 3, pp. 377-387.

13. Caballero J., Song D. Automatic protocol reverse-engineering: Message format extraction and field semantics inference. *Computer Networks*, 2013, vol. 57, no. 2, pp. 451-474.

14. Caballero J., Poosankam P., Kreibich C., Song D. Dispatcher: Enabling active botnet infiltration using automatic protocol reverse-engineering. *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009, pp. 621-634.

15. Wang Z., Jiang X., Cui W., Wang X., Grace M. ReFormat: Automatic reverse engineering of encrypted messages. *European Symposium on Research in Computer Security*. Springer, Berlin, Heidelberg, 2009, pp. 200-215.

16. Lin Z., Zhang X., Xu D. Reverse engineering input syntactic structure from program execution and its applications. *IEEE Transactions on Software Engineering*, 2009, vol. 36, no. 5, pp. 688-703.

17. Narayan J., Shukla S. K., Clancy T. C. A survey of automatic protocol reverse engineering tools. *ACM Computing Surveys*, 2016, vol. 48, no. 3, pp. 40:1-40:26.

18. Wang C., Li S. CoRankBayes: Bayesian learning to rank under the co-training framework and its application in keyphrase extraction. *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 2241-2244.

19. Isa D., Lee L. H., Kallimani V. P., Rajkumar R. Text document preprocessing with the Bayes formula for classification using the support vector machine. *IEEE Transactions on Knowledge and Data engineering*, 2008, vol. 20, no. 9, pp. 1264-1272.

20. Jiang X., Hu Y., Li H. A ranking approach to keyphrase extraction. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 756-757.

21. Lopez P., Romary L. HUMB: Automatic key term extraction from scientific articles in GROBID. *Proceedings of the 5th international workshop on semantic evaluation*. Association for Computational Linguistics, 2010, pp. 248-251.

22. Frazão. I., Abreu P., Cruz T. Araujo A., Simoes P. Cyber-security Modbus ICS dataset. *IEEE Dataport*, 2019. DOI: 10.21227/pjff-1a03.

Статья поступила 15 декабря 2019 г.

Информация об авторах

Гайфулина Диана Альбертовна – младший научный сотрудник лаборатории кибербезопасности и постквантовых вычислений. Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: информационная безопасность, киберфизические системы, интеллектуальный анализ данных, обнаружение аномалий в среде передачи данных. E–mail: gaifulina@comsec.spb.ru

Котенко Игорь Витальевич – доктор технических наук, профессор. Заведующий лабораторией проблем компьютерной безопасности. Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: безопасность компьютерных сетей, искусственный интеллект, телекоммуникационные системы. E–mail: ivkote@comsec.spb.ru

Федорченко Андрей Владимирович – младший научный сотрудник лаборатории проблем компьютерной безопасности. Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: безопасность компьютерных сетей, обнаружение вторжений, вредоносные программы. E–mail: fedorchenko@comsec.spb.ru

Адрес: 199178, Россия, г. Санкт-Петербург, 14 линия, д. 39.

A Technique for Lexical Markup of Structured Binary Data for Problems of Protocols Analysis in Uncertainty Conditions

D. A. Gaifulina, I. V. Kotenko, A. V. Fedorchenko

Purpose. *The study of network protocols for complex information and communication infrastructures, such as cyberphysical systems or the Internet of Things, is subject to decrease due to the uncertainty in the structure of the data stored and transmitted in them. Thus, there is a need to define specifications for such unregulated protocols. The initial data in this case is binary data containing protocol messages, for example, network traffic, in the form of a sequence of atomic information units (bytes, bits).**The purpose** of the work is to overcome the network traffic lexical uncertainty based on marking up traffic by message tokens (fields) and forming common lexical structures (protocols). An approach is proposed to identify typical structures of protocols for generating messages in binary data and to determine their lexical specifications.* ***Methods.*** *The approach is based on statistical methods of analyzing information in natural languages, which are widely used in the intellectual analysis of texts, which is the main theoretical contribution of this work.* ***Novelty.*** *The novelty of the proposed approach lies in the possibility of lexical recognition of conditionally structured binary data based on the frequency analysis of possible sequences of information units and their combinations.* ***Results.*** *As a result of the research, a methodology for the analysis of structured binary data was developed. The approach to the analysis of binary data was tested on various datasets of network traffic under the conditions of uncertain specifications of network protocols presented in the form of an ordered sequence of digital (binary) data. According to the results of an experimental assessment of the accuracy of the developed methodology, the identified typical sequences (tokens) of information units (bytes) on average correspond to the content of the fields of network traffic protocols by 95%, and the identified structures (sequences of tokens) correspond to the specifications of these protocols by 70%.* ***Practical relevance.*** *The ability to perform pre-processing of structured binary data to automatically solve problems of detecting anomalies, perform a security audit, identify protocol vulnerabilities, conduct application testing, and investigate information security incidents in the face of undefined specifications of interaction protocols in the face of uncertainty.*

Key words: *recovery of data formats, structural analysis, binary data, proprietary protocols, security analysis, uncertain infrastructures*

Information about Authors

Diana Albertovna Gaifylina – Junior Research Associate of Laboratory of Cybersecurity and Post-quantum Cryptography. St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science. Field of research: information security, cyber-physical systems, data mining, detection of anomalies in the data transfer medium. E-mail: gaifulina@comsec.spb.ru

Igor Vitalievich Kotenko – Dr. habil. of Engineering Sciences, Professor. Head of Laboratory of Computer Security Problems. St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science. Field of research: information security, artificial intelligence, telecommunications. E-mail: ivkote@comsec.spb.ru

Andrey Vladimirovich Fedorchenko – Junior Research Associate of Laboratory of Computer Security Problems. St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science. Field of research: computer network security, intrusion detection, malware. E-mail: fedorchenko@comsec.spb.ru

Address: Russia 199178, Saint-Petersburg, 14th Liniya, 39.