프로토콜 리버스 엔지니어링의 정교한 정적필드 추출 방법 제안

이민섭 º, 심규석, 구영훈, 김명섭

고려대학교 컴퓨터정보학과

{chenlima2, kusuk007, gyh0808, tmskim}@korea.ac.kr

요 약

오늘날의 네트워크 환경은 매우 급속도로 성장하고 있으며, 이로 인해 인터넷 트래픽이 기하급수적으로 증가하고 있다. 그 결과로 여러 응용 및 악성 행위도 급증하여 복잡하고 다양한 비공개 프로토콜이 발생하고 있다. 이러한 비공개 프로토콜들의 구조를 분석하는 프로토콜 리버스 엔지니어링은 네트워크 관리 및 보안 분야에서 필수적인 요소로 자리잡고 있다. 기존의 다양한 연구에서 프로토콜 리버스 엔지니어링을 다뤄왔지만 메시지 내에 필드들을 구분하거나 추출하는 표준화된 방법은 없다. 따라서 본 논문에서는 정교한 정적필드 추출 방법을 제안한다.

1. 서 론

오늘날의 네트워크 환경은 매우 급속도로 성장하고 있으며, 이로 인해 인터넷 트래픽이 기하급수적으로 증가하고 있다. 그 결과로 여러 응용 및 다양한 악성 행위들이 나타나고 있다. 이러한 환경에서 발생하는 복잡하고 다양한 프로토콜들은 보통 알려지지 않은 비공개 프로토콜이다. 비공개 프로토콜의 구조를 분석하는 것은 프로토콜의 메시지 형식과 의미, 순서 같은 상세한 구조를 추출하는 것을 목표로 한다. 따라서 프로토콜 리버스 엔지니어링은 네트워크 관리 및 보안 분야에서 필수적인 요소로 자리잡고 있다. 네트워크 관리 분야 에서는 프로토콜별 네트워크 사용 현황 파악, 한정되어 있는 네트워크 자원을 효율적으로 사용하기 위해 특정 프로토콜에 대한 대역폭 조절 등 네트워크 관리에 활용이 가능하다. 네트워크 보안 분야에서는 매년 꾸준히 급증하는 여러 악성 행위들이 발생시키는 비공개 프로토콜의 구조를 분석함으로써 특정 악성 행위에 대한 정보를 습득하여 대처하거나 기존에 알려져 있지 않은 공격에 대한 탐지 시스템을 구축하는데 도움이 될 수 있다.

기존의 다양한 연구에서 프로토콜 리버스 엔지니어링을 다뤄왔지만, 현재까지 표준화된 필드 구분 및 추출 방법은 존재하지 않으며 각 연구마다 각각의 장단점이 존재한다. 따라서 본 논문에서는 여러 연구들의 장점을 결합하여 프로토콜 리버스 엔지니어링에서의 정교한 정적필드 추출 방법을 제안한다.

이 논문은 2015 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No.2015R1D1A3A01018057)과 2017 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No. 2017-0-00513, Security Analytics 기반의 이기종 보안솔루션 위협 분석 및 대응 기술 개발)

2. 본 론

1) 비공개 프로토콜 구조분석 구성 요소

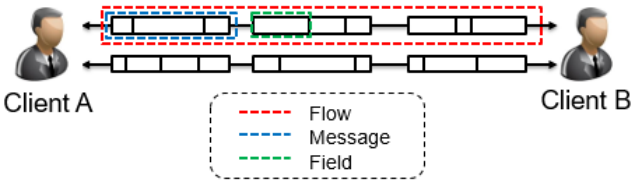

Figure 1. 비공개 프로토콜 구조분석 구성 요소

본 논문에서 기술하는 필드를 정의하기 위해서 프로토콜 리버스 엔지니어링의 입력으로 사용되는 플로우(Flow)와 플로우를 구성하는 메시지(Message)를 먼저 정의한다. 플로우란, 메시지의 연속이다. 일반적으로 메시지의 단위는 TCP 플로우의 경우, 하나의 TCP 세그먼트로 하고 UDP 의 경우, 하나의 패킷으로 한다. 메시지는 필드 단위로 세분화 될 수 있다. 즉, 메시지는 필드의 연속이다. 필드는 프로토콜의 구조 분석에서 의미를 가지는 가장 작은 단위이다. 필드는 길이, 값, 값의 구조에 따라 분류되는데 본 논문에서는 값이 고정적인 정적 필드(static field)를 추출하는 방법에 대해 기술한다.

2) 정교한 정적필드 추출 방법 제안

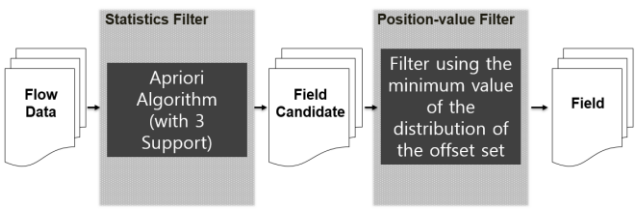

Figure 2. 정교한 정적필드 추출 방법 Overview

본 논문에서는 필드를 추출하기 위해 두가지 단계를 정의한다. 첫 번째로 패턴 마이닝의 한 종류인 Apriori 알고리즘에 세가지 지지도를 적용하는 Statistics Filter 단계, 두 번째로 필드의 메시지 내의 위치값들의 분포를 고려하는 Position-value Filter 단계가 존재한다.

$$flow_{supp} = \frac{number\ of\ flows\ containing\ item}{total\ number\ of\ flows}$$

$$message_{supp} = \frac{number\ of\ messages\ containing\ item}{total\ number\ of\ messages}$$

$$flow\_set_{supp} = \frac{number\ of\ flow\_set\ containing\ item}{total\ number\ of\ flow\_set}$$

Figure 3. 3 가지 지지도 정의

첫 번째 단계로 Statistics Filter 단계에서는 모든 1 바이트 Character 를 입력으로 받아서 Apriori 알고리즘에 Figure 3 의 세가지 지지도($flow_{supp}$, $message_{supp}$, $flow\_set_{supp}$)를 사용하여 k 길이의 빈번한 문자열 집합을 추출한다. $flow_{supp}$ 는 해당 문자열이 플로우 내에서 얼마나 빈번하게 발생하는지에 대한 지표로써 메시지에서는 빈번하게 발생하지만 몇몇 플로우에서만 발생하는 문자열들을 제거할 수 있다. $message_{supp}$ 는 해당 문자열이 메시지 내에서 얼마나 빈번하게 발생하는지에 대한 지표로써 Statistics Filter 단계에서 필드후보들을 추출하는 핵심적인 지지도이다. $flow\_set_{supp}$ 에서 $flow\_set$ 이란 하나의 서버와 해당 서버랑 통신하는 클라이언트 간에 연결을 맺고있는 모든 플로우들을 원소로 하는 집합을 의미한다. 즉, $flow\_set_{supp}$ 는 해당 문자열이 플로우 집합 내에서 얼마나 빈번하게 발생하는지에 대한 지표로써 몇몇 플로우에서만 빈번하게 출현하는 문자열들을 제거할 수 있다.

Statistics Filter 단계에서는 k 길이의 문자열 마다 위 3 가지 지지도를 구하고 각각 지지도 모두 최소 지지도를 만족하는 k 길이의 빈번한 문자열들을 필드후보로 추출한다.

Figure 4. Startoffset 과 Endoffset 정의

두 번째 단계로 Position-value Filter 단계에서는 Figure 5 와 같이 Statistics Filter 단계를 거쳐서 만들어진 k 길이의 빈번한 문자열(i_k)의 메시지 시작부분 기준 Offset(Startoffset)과 메시지 끝부분 기준 Offset(Endoffset)의 분산을 고려하여 메시지의 특정 위치에 비교적 고정적으로 나타나는 문자열들을 최종 필드로 추출한다.

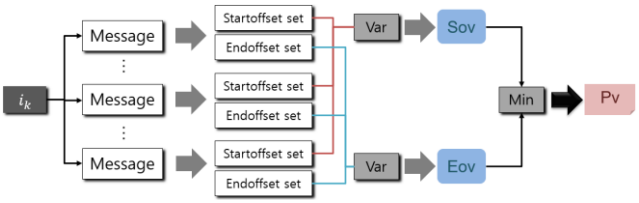

Figure 5. Position-value Filter 요약

필드 후보들이 각 메시지 마다 어느 위치에 나타나는지를 의미하는 Startoffset 과 Endoffset 을 계산한다. 그리고 필드후보 한 개와 해당 필드후보가 나타나는 메시지 한 개당 Startoffset 의 집합인 So(Startoffset set)와 Endoffset 의 집합인 Eo(Endoffset set)를 구성한다. 그러면 한 개의 필드후보에 필드후보가 나타나는 메시지 개수만큼의 So 와 Eo 가 생성되는데 이 So 들의 분산(Sov)과 Eo 들의 분산(Eov)을 구한다. 그리고 Sov 와 Eov 중에 최솟값(Pv)을 선택한다. Pv 는 특정 필드후보가 메시지 내에서 어떻게 분포되어 있는지를 나타내는 지표이다. Pv 가 작으면 작을수록 해당 필드 후보는 고정적인 위치에 나타난다는 것을 의미한다. 따라서 Pv 가 특정 Threshold 를 만족하면 이 문자열은 필드 집합에 저장한다.

Position-value Filter 단계를 거치면 메시지 내 필드들의 집합이 생성된다. 이 필드들은 메시지 내에서 빈번하게 발생하면서 어느정도 고정적인 위치에 발생하는 필드로 정의된다.

3. 결론 및 향후 연구

기존의 프로토콜 리버스 엔지니어링을 다뤘던 논문들은 이미 알려져 있는 구분자로 필드를 구분하거나 통계적 정보 혹은 위치적 정보 중 한가지 방법만으로 메시지 내 필드를 구분하고 있다. 이러한 방법들은 필드의 한가지 정보만을 가지고 필드를 구분하기 때문에 잘못된 필드가 추출된다는 한계점이 있다. 하지만 본 논문에서는 필드의 통계적 정보와 위치적 정보를 동시에 고려하여 필드를 추출하기 때문에 플로우와 메시지에서 빈번하게 발생하면서 고정적인 위치에 나타나는 정교한 필드를 추출할 수 있다. 향후 연구로는 값이 고정적인 정적필드를 제외한 다른 필드들을 효율적으로 추출할 수 있는 방법을 연구할 계획이다.

4. 참고 문헌

[1] 구영훈, Baraka D. Sija, 김명섭, “프로토콜 리버스 엔지니어링의 이상적인 메커니즘 정의”, 통신망 운용관리 학술대회(KNOM 2017),pp. 23-24, 2017 년 6 월.

[2] Jian-Zhen Luo and Shun-Zheng Yu, Position-based automatic reverse engineering of network protocols, 2013.