네트워크 트레이스 기반 프로토콜 리버스 엔지니어링 시스템 아키텍쳐

구영훈, 심규석, 박지태, Huru Hasanova, 김명섭
고려대학교
{gyh0808, kusuk007, pjj5846, hhuru, tmskim}@korea.ac.kr

Architecture of Network Trace-based Protocol Reverse Engineering System

Young-Hoon Goo, Kyu-Seok Shim, Jee-Tae Park, Huru Hasanova, Myung-Sup Kim
Korea Univ.

요 약

오늘날 급격하게 발전하는 인터넷 환경 하에 복잡하고 다양한 비공개 프로토콜의 트래픽이 발생하고 있다. 프로토콜 리버스 엔지니어링은 비공개 프로토콜의 사양을 추출하는 작업이며 많은 네트워크 관리 및 보안 문제를 해결하는 데에 유용하다. 이를 위해 다양한 프로토콜 리버스 엔지니어링 시스템이 제안되었지만, 실제 네트워크 관리에 활용하기에는 많은 한계점들이 존재한다. 네트워크 트레이스 기반 프로토콜 리버스 엔지니어링은 실행 트레이스 기반 프로토콜 리버스 엔지니어링보다 실용적이지만, 이 역시 프로토콜의 메시지 페이로드 내용을 분석하기 때문에 시간이 오래 걸리며, 수집한 트래픽의 환경에 따라 의존성이 있는 잘못된 사양이 추출될 수 있는 오류가 발생하기 쉽다. 본 논문에서는 지속적인 유지보수를 통해 정확성 측면에서 강건한 프로토콜 사양을 추출할 수 있는 네트워크 트레이스 기반 프로토콜 리버스 엔지니어링 시스템의 아키텍쳐에 대해 설명한다.

Ⅰ. 서론

오늘날 급진적으로 발전하는 초고속 인터넷 환경에서는 대용량의 트래픽이 발생하고 있으며, 이에 따라 다양한 기능의 응용과 악성 행위가 기하급수적으로 증가하고 있다. 이러한 환경 하에 발생하는 복잡 다양한 트래픽 중 다수는 대부분 알려지지 않았거나 최소한으로 문서화되어 있는 독점적인 프로토콜이다. 대표적으로 Skype 프로토콜, 안티바이러스 도구가 사용하는 소프트웨어 업데이트 프로토콜, SCADA(Supervisory Control and Data Acquisition) 시스템의 프로토콜, Botnet의 C&C(Command and Control) 프로토콜 등이 이러한 비공개 프로토콜이라 할 수 있다. 비공개 네트워크 프로토콜의 사양을 추출하는 작업인 프로토콜 리버스 엔지니어링은 많은 네트워크 관리 및 보안 문제를 해결하는 데에 매우 유용하다. 예를 들어, 네트워크 모니터링 분야에서는 대상 네트워크에서 발생하는 알 수 없는 트래픽의 정보를 습득할 수 있으므로, 비공개 프로토콜이 발생시키는 트래픽을 분류하여 네트워크 사용 현황 파악, 확장 계획 수립 및 QoS(Quality of Service) 정책 설정에 활용이 가능하다. 네트워크 보안 분야에서는 이전에 알려지지 않은 공격의 탐지 및 차단을 위한 방화벽과 침입 탐지 시스템에 도움이 될 수 있으며 네트워크 취약성을 파악하기 위한 침투 시험 및 스마트 Fuzzer 시스템 구축에 유용한 정보를 제공할 수 있다. 또한 레거시 소프트웨어 통합, 지능형 DPI에도 활용이 가능하다.

이를 위해 다양한 프로토콜 리버스 엔지니어링 시스템이 제안되었지만, 실제 네트워크 관리에 활용하기에는 많은 한계점들이 존재한다. 네트워크 트레이스 기반 프로토콜 리버스 엔지니어링은 실행 트레이스 기반 프로토콜 리버스 엔지니어링보다 실용적이지만[1], 이 역시 프로토콜의 메시지를 수집하고 실제 페이로드 내용을 분석하기 때문에 시간이 오래 걸리며, 수집한 트래픽의 환경에 따라 의존성이 있는 잘못된 사양이 추출될 수 있는 오류가 발생하기 쉽다. 본 논문에서는 지속적인 유지보수를 통해 정확성 측면에서 강건한 프로토콜 사양을 추출할 수 있는 네트워크 트레이스 기반 프로토콜 리버스 엔지니어링 시스템의 아키텍쳐에 대해 설명한다.

본 논문의 구성은 본 장의 서론에 이어 2장에서 해결하고자 하는 문제를 정의하고 3장에서 제안하는 프로토콜 리버스 엔지니어링 시스템의 아키텍쳐에 대해 기술한다. 4장에서는 결론 및 향후 연구에 대해 기술한다.

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2015R1D1A3A01018057).

Ⅱ. 문제 정의

두 장치 간에 이루어지는 통신은 일련의 프로토콜을 필요로 한다. 프로토콜은 데이터 통신을 위한 규칙의 집합이다. 즉, 두 호스트 간 통신 시 교환되는 메시지 유형들과 이를 구성하는 필드 형식은 물론, 이러한 메시지의 송신 또는 수신시 수행되는 순서를 정의한다. 프로토콜 리버스 엔지니어링은 분석하고자 하는 프로토콜 사양이 어떠한 유형들의 메시지들을 갖고 있는지, 이러한 메시지들이 어떠한 순서로 동작하는지를 추론해야 한다. 이전의 많은 연구에서는 메시지 유형들의 동작 순서를 언어의 문법을 모델링할 수 있는 유한 상태 머신(FSM)으로 표현하여 추출하였다. 따라서, 프로토콜 리버스 엔지니어링의 결과로 추출되는 프로토콜 사양은 다음과 같이 정의할 수 있다.

$$Specification = \{ \Sigma\ Message\ Type = \{ M_1, M_2, \cdots, M_n \}, \Sigma\ Transition = \{ T_1, T_2, \cdots, T_k \} \} \quad (1)$$

$$M = \{ F_1, F_2, \cdots, F_m \} \quad (2)$$

$$T = \{ transition\ probability, M_i \rightarrow M_j \} \quad (3)$$

수식 (1)은 프로토콜의 사양을 의미하며, 메시지 유형들과 메시지 유형 간 전이들을 원소로 갖고 있다. 수식 (2)는 하나의 메시지 유형을 의미하며, 구성하는 필드들을 순서대로 원소로 갖고 있다. 수식 (3)은 하나의 메시지 유형 간 전이를 의미하며, 이에 해당하는 전이 확률을 원소로 갖고 있다. 이를 통해, 프로토콜의 각 메시지 유형의 형식(2) 표현과 각 메시지 유형(2)을 state로 하고, 각 메시지 유형 간 전이(3)를 transition으로 하는 프로토콜 FSM 표현이 가능하다.

기존의 많은 연구들은 메시지 유형(M)의 형식을 추출하기 위하여 Zipf 법칙과 샤논 이론의 Entropy H(x), 연관 규칙 마이닝의 Support, LDA의 출현 확률과 같이 특정 키워드의 빈도를 기반으로 하는 알고리즘을 사용

하여 필드 F를 추출한다. 이를 수행하기 위해서는 프로토콜의 트래픽에서 메시지들의 페이로드의 내용을 가지고 분석하여야 하며, 이는 큰 시간복잡도로 이어진다. 따라서, 추출한 비공개 프로토콜의 사양을 관리하여 재사용할 필요가 있다.

한편, 트래픽을 추출한 네트워크 환경이 다양하지 못하면 특정 환경에 의존성이 있는 잘못된 키워드가 필드 형식(F)으로 추출될 수 있다. 예를 들어, 특정 URL이나 HTTP 프로토콜 Header Line의 Date : 필드의 특정 시간 값, Host : 필드의 특정 호스트 정보, User-Agent : 필드의 특정 웹 브라우저 환경이 필드 형식(F)로 추출된다면 잘못된 메시지 유형(M) 및 FSM 추출을 야기한다. 따라서, 프로토콜 사양 관리 DB에서 재사용을 위한 관리와 동시에 정확한 프로토콜 사양으로 갱신을 하기 위한 보수 작업이 필요하다.

또한, 네트워크 트레이스 기반 프로토콜 리버스 엔지니어링을 위해서는 반드시 입력이 다중의 프로토콜이 아닌, 단일의 비공개 프로토콜에 대한 트래픽 트레이스이어야 한다는 전제 조건이 필요하다.

Ⅲ. 시스템 아키텍처

본 장에서는 네트워크 트레이스 기반 프로토콜 리버스 엔지니어링 시스템의 아키텍처에 대해 설명한다.

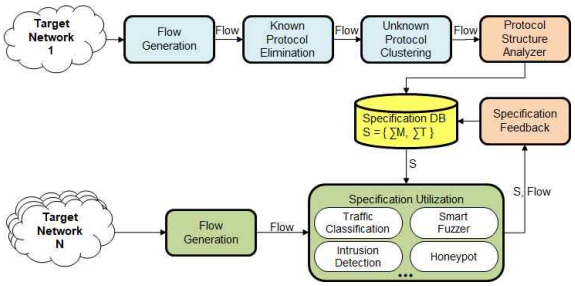

그림 1. 제안하는 프로토콜 리버스 엔지니어링 시스템 아키텍처

그림 1은 본 논문에서 제안하는 네트워크 기반 프로토콜 리버스 엔지니어링 시스템의 아키텍처를 나타낸다. 본 아키텍처는 신속한 트러블 슈팅을 위해 다양한 비공개 프로토콜 트래픽 발생 네트워크에서 생성한 사양을 중앙에 위치한 Specification DB에서 통합 관리하며, 프로토콜 사양이 필요한 분석 네트워크에 배포한다. 또한, 초기 네트워크에서 생성한 비공개 프로토콜 사양(S)을 더욱 정확한 사양으로 만들기 위해 다양한 네트워크에서의 분석 이력을 기반으로 해당 사양(S)을 갱신한다. 아키텍처는 크게 3 부분으로 구성된다. 단일 비공개 프로토콜의 트래픽만을 수집하기 위한 단일 프로토콜 수집부, 프로토콜 사양을 추출하고 관리 및 유지 보수를 위한 프로토콜 분석・관리부, 그리고 추출한 프로토콜 사양을 다양한 네트워크 관리 및 보안 목적으로 활용하기 위한 프로토콜 사양 활용부이다.

단일 비공개 프로토콜 수집부는 초기 네트워크에서 수집한 트래픽을 5-tuple(SrcIP, SrcPort, Dst IP, DstPort, Transport Layer Protocol)이 같은 패킷의 집합인 플로우 단위로 변경(Flow Generation)하고 선행적으로 다량의 알려진 프로토콜의 트래픽을 분류하여 필터링(Known Protocol Elimination)한다. 이미 알려진 프로토콜의 트래픽은 헤더 기반, 페이로드 기반, 통계 기반, 행위 기반 시그니쳐 및 상관 관계 분석 등의 방법을 통해 분류할 수 있다[2]. 다음으로 필터링 후 분류되지 않은 소량의 트래픽에서 단일 비공개 프로토콜 단위의 트래픽을 수집하기 위하여 유사한 통계 정보 패턴을 가지는 트래픽들을 그룹핑(Unknown Protocol Clustering)한다. 단일 비공개 프로토콜의 트래픽을 클러스터링하는 방법으로는 [3]의 플로우의 연관성 모델을 이용하거나 [4]의 플로우의 다양한 특성 값을 활용하여 기계 학습을 통해 수행하는 방법이 있다.

프로토콜 분석・관리부에서는 단일 프로토콜 수집부에서 트래픽을 전달받아 프로토콜 사양(S)을 추출(Protocol Structure Analyzer)한다. 추출된 사양(S)은 Specification DB에 저장된다. 그리고, 프로토콜 사양 활용부에서 활용된 사양(S)의 사용 이력을 기반으로 더 정확한 사양으로 갱신(Specification Feedback)한다. 프로토콜 사양(S) 갱신은 각 메시지 유형 형식(M)이 매칭된 트래픽의 양(flow, packet, byte), 분석에 사용되지 않은 횟수, 분석한 트래픽의 고유한 클라이언트의 수의 비율 등과 같은 다양한 척도들을 이용하여 각 메시지 유형 형식(M)에 Weight를 설정하고 Weight가 일정 임계값 이하일 시 제거하는 방법을 사용할 수 있다. 또한, 메시지 형식(M)을 구성하는 필드 형식(F)들 중 일부분만 매칭이 된 트래픽들만을 모아 프로토콜 사양을 재추출하고 원본 메시지 형식(M)과 비교하여 잘못된 원본 메시지 유형 형식(M)을 올바른 메시지 형식으로 갱신할 수 있다. 메시지 유형 간 전이(T)의 확률(Transition Probability) 갱신 역시, 분석한 트래픽과 사용된 메시지 유형(M) 기반으로 이루어지며 이는 Honeypot 시스템에서 더 정확한 패킷 재생에 활용될 수 있다.

프로토콜 사양 활용부는 대상 네트워크의 트래픽을 플로우 단위로 변경(Flow Generation)하고, 추출한 사양(S)를 활용하여 다양한 네트워크 관리 및 보안 목적에 활용(Specification Utilization)한다. 활용한 사양(S)과 트래픽 및 분석 결과는 프로토콜 사양 분석・관리부에 전달하여 더 정확한 사양으로 갱신한다.

Ⅳ. 결론 및 향후 연구

프로토콜 리버스 엔지니어링은 효율적인 네트워크 관리 및 보안을 위해 유용한 정보를 제공할 수 있다. 이를 위해 다양한 방법론이 제안되었지만, 실제 네트워크에 적용하기에 많은 한계점을 가지고 있다. 본 논문에서는 정확한 프로토콜 사양 추출 및 관리를 위한 네트워크 트레이스 기반 프로토콜 리버스 엔지니어링 시스템의 아키텍처를 제안하였다. 향후 연구로는 제안한 아키텍처를 기반으로 프로토콜 리버스 엔지니어링 시스템을 개발하여 실제 네트워크에 적용할 계획이다.

참 고 문 헌

[1] I. Bernudez, A. Tongaonkar, M. Iliofotou, M. Mellia, and M. M. Munafo, “Automatic Protocol Field Inference for Deeper Protocol Understanding”, Proc. of the 14th IFIP Networking Conference, Toulouse, France, May. 2015.

[2] S.-H. Yoon, K.-S. Shim, S.-K. Lee, and M.-S. Kim, "Framework for Multi-Level Application Traffic Identification," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2015, Busan, Korea, Aug. 2015, pp.424-427.

[3] Y.-H. Goo, S.-H. Lee, K.-S. Shim, B. D. Sija, and M.-S. Kim, “A Traffic-Classification Method Using the Correlation of the Network Flow”, J. KIISE. Vol.44, No.4, pp.433-438. April. 2017.

[4] J.-T. Park, K.-S. Shim, S.-H. Lee, and M.-S. Kim, “FloFex : Traffic Learning Feature Extraction System for Accurate Application Traffic Classification based on Machine Learning”, Proc. of KNOM 2017 Gwangu, Korea,J. June. 2017.