클러스터링 기반 비트코인 랜섬웨어 주소의 Ground-Truth 생성 방법

김보선, 백의준, 지세현, 강민규, 신희종, 김명섭
고려대학교 컴퓨터정보학과
{boseon12, pb1069, sxzer, cxz3619, tlswhd0215, tmskim}@korea.ac.kr

A Method for Ground-Truth Generation of Bitcoin Ransomware Addresses Based on Clustering

Boseon Kim, Ui-Jun Baek, Se-Hyun Ji, Mingyu Gang, Huijong Sin, Myung-Sup Kim
Computer and Information Science, Korea University

요 약

추적이 불가능한 거래인 비트코인이 등장하며, 익명성과 낮은 거래비용과 같은 비트코인의 특징을 악용하는 사례가 증가하고 있다. 대표적으로 랜섬웨어는 다양한 양상으로 진화하고 있다. 본 논문은 비트코인 트랜잭션을 수집하여 K-means 클러스터링을 통해 정상적인 주소와 알려진 랜섬웨어 주소를 구분하고자 한다. 이를 통해 랜섬웨어 주소가 포함되지 않은 클러스터와 랜섬웨어 주소 포함 비율이 높은 클러스터를 알 수 있었다.

Ⅰ. 서 론

랜섬웨어는 특정 시스템에 침투 후 사용자의 파일을 암호화하여 사용자가 열람할 수 없도록 하고 복구 조건으로 몸값(Ransom)을 요구하는 악성 소프트웨어다[1]. 초기에는 낮은 PC 보급률로 인해 피해 규모가 작았지만, 현재는 PC 보급률이 증가하며 피해 규모도 커지고 있다.

IT 기술이 발전함에 따라 추적이 불가능한 거래인 비트코인이 등장하였다. 비트코인은 2009년 사토시 나카모토라는 가명을 쓰는 프로그래머에 의해 개발된 암호화폐로 중앙은행 없이 전 세계에서 P2P 거래 방식으로 개인 간에 자유로운 금융거래를 할 수 있으며, 익명성과 낮은 거래비용으로 기존의 어떠한 화폐보다 사용하기 편리하다[2]. 이로 인해 랜섬웨어 공격자가 직접적으로 금전적 이윤을 얻을 수 있는 구조가 갖춰져 랜섬웨어의 규모가 기하급수적으로 증가하게 되었다.

WannaCry는 2017년 5월에 등장한 랜섬웨어로 인터넷 네트워크에 접속만 해도 컴퓨터를 감염시켜 암호화된 파일을 푸는 대가로 비트코인을 요구하는 메시지를 띄웠다. 이처럼 랜섬웨어에 의한 경제적 손실을 막기 위해 다양한 랜섬웨어 예방 및 탐지 시스템에 대한 연구가 진행되고 있지만, 랜섬웨어를 완벽하게 차단하기는 어렵다.

비트코인은 모든 거래내역이 공개되어 있어서 누구나 분석이 가능하다. 비트코인 주소를 통해 사용자를 식별할 수 없지만 대부분 비트코인을 환전하기 위해 중개 거래소를 이용하여 본인 휴대전화나 계좌 인증이 필요하다. 클러스터링을 통해 어떤 주소와 비슷한 특징을 가진 주소들을 묶을 수 있다. 랜섬웨어에 관련된 모든 주소가 알려지지 않았기 때문에 Kaggle에서 수집한 281,638개의 랜섬웨어 주소를 제외하고 본 연구팀이 수집한 2,484,978개의 비트코인 주소를 정상 주소라고 판단할 수 없다. 따라서 랜섬웨어 주소가 포함되지 않은 클러스터를 정상 주소로 판단하여 정답지로 활용할 수 있으며, 랜섬웨어에 관련된 주소 포함 비율이 높은 클러스터의 나머지 주소 또한 랜섬웨어 주소라고 의심할 수 있으므로 정상 주소와 랜섬웨어 관련 주소를 분류해야 한다. 본 논문에서는 클러스터링을 통해 정상 주소와 랜섬웨어 관련 주소를 분류하고자 한다.

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2018R1D1A1B07045742)과 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2018-0-00539-003,블록체인의 트랜잭션 모니터링 및 분석 기술개발)

Ⅱ. 본론

본 장에서는 데이터 수집과 랜섬웨어 주소 분석 방법에 대해 언급한다.

A. 비트코인 블록 및 트랜잭션 데이터 수집 및 전처리

실험에 사용된 데이터는 2009년부터 2011년 비트코인 블록 데이터와 Kaggle에서 수집한 2009년부터 2011년 랜섬웨어 주소 데이터 셋을 사용한다. 수집한 트랜잭션 기준 데이터를 주소 기준 데이터로 변환한다.

B. 특징 추출

주소 기준 데이터에서 추출한 21가지 특징을 설명하며, 표 1과 같다.

표 1 주소 데이터에서 추출한 특징 의미

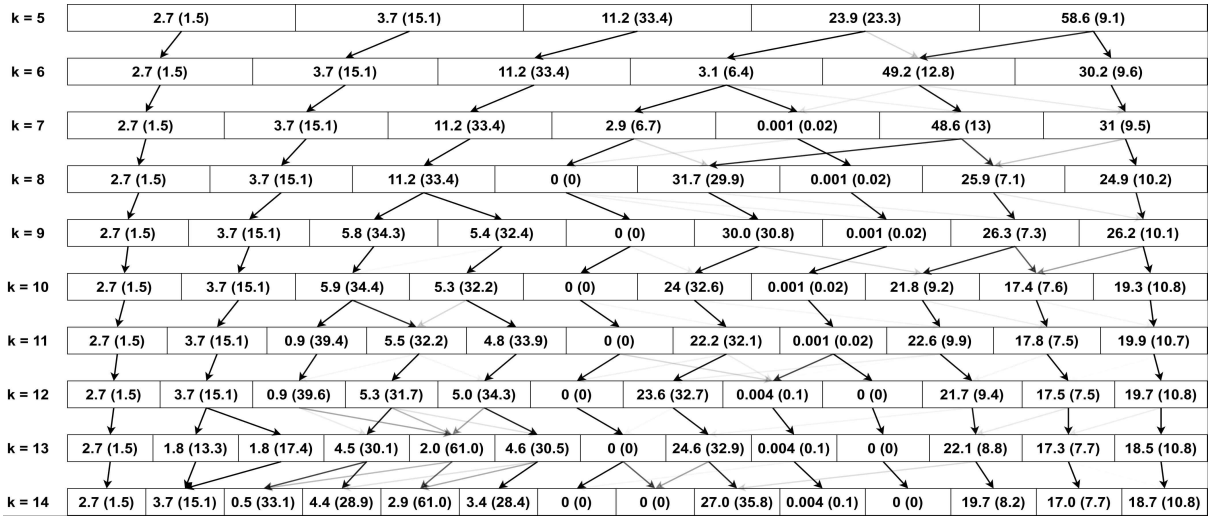
항목	항목	의미
all_	count	전체 거래 횟수
	amount	전체 거래 금액
	amount_mean	전체 거래 금액 평균
	amount_std	전체 거래 금액 표준편차
	activation_period	전체 거래 시간
	activation_mean	전체 거래 시간 평균
	activation__std	전체 거래 시간 표준편차
send_	count	보낸 거래 횟수
	amount	보낸 거래 금액
	…	…
recv_	count	받은 거래 횟수
	amount	받은 거래 금액
	…	…

C. K-means Clustering

K-means 알고리즘은 가장 일반적으로 사용되는 클러스터링 알고리즘이다[3]. K-means 알고리즘은 데이터를 k 개의 클러스터로 묶는 알고리즘이다. 다음 수식 1과 같다.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \qquad (1)$$

그림 1 클러스터 개수별 랜섬웨어 주소 포함 비율



Ⅱ-B 절에서 추출한 21가지 특징을 기준으로 K-means 클러스터링을 진행하였다. 클러스터링 결과는 그림 1과 같다. 클러스터 개수(k)를 5부터 14로 1씩 증가하여 실험하였다. 한 행은 클러스터 개수(5,6,7,8,9,10,11,12,13,14)를 의미하며, 사각형 하나는 클러스터 하나를 나타낸다. 화살표는 k가 n 일 때 a 번째 클러스터(a<=n)의 주소가 k가 n+1일 때 b 번째 클러스터(b<=n+1)로 이동하는 것을 의미한다. 사각형 안에 첫 번째 숫자는 전체 랜섬웨어 주소 중 클러스터 내에 랜섬웨어 주소 포함 비율이며, 괄호 안에 든 숫자는 클러스터 내에 전체 주소 중 랜섬웨어 주소 포함 비율이다. 예를 들어 k가 6일 때 4번째 클러스터는 k가 7일 때 4번째와 5번재 클러스터로 나누어진다. 해당 클러스터의 주소가 클러스터 개수가 증가함에 따라 2개의 클러스터로 분할되었음을 의미한다. 실험을 통해 클러스터 내 전체 주소 중 높은 비율의 랜섬웨어 주소가 포함된 클러스터를 확인하였다. k가 14일 때 5번째 클러스터는 클러스터 내 주소(13,221개) 중 61%(8,042개)가 랜섬웨어 주소인 것을 알 수 있었다. 또한 k가 8부터 14까지 랜섬웨어 주소가 속하지 않은 클러스터를 찾을 수 있었다. k가 8일 때 4번째 클러스터는 0(0)인 것을 보아 랜섬웨어 주소를 한 개도 포함하고 있지 않으며, 해당 클러스터에 77,827개의 주소가 있다. k가 14일 때 7, 8, 11번째 클러스터도 랜섬웨어의 주소를 갖고 있지 않으며 총 96,827개의 주소가 있다. 클러스터의 개수가 증가하며 랜섬웨어와 확실하게 구분되는 주소의 개수가 증가하였다. k가 14일 때 10번째 클러스터 또한 해당 클러스터의 전체 주소 중 해당 클러스터의 랜섬웨어 주소 포함 비율이 0.1%도 되지 않으며, 클러스터 내에 12,980개의 주소가 있다. 이를 통해 k가 14일 때 7, 8, 10, 11번째 클러스터의 109,807개의 주소는 정상적인 주소라고 판단할 수 있다.

D. Silhouette Coefficient

본 절에서는 클러스터 개수별 Silhouette score를 나타낸다. 실루엣 계수는 군집화의 성능을 판단하기 위한 기준이다. 모든 데이터 에 대해 a 는 i와 같은 군집에 속한 원소들의 평균 거리이며, b_i는 i와 다른 군집 중 가장 가까운 군집까지의 평균 거리이다. 수식은 다음과 같다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1 \qquad (2)$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases} \qquad (3)$$

표 2 클러스터 별 실루엣 계수

#클러스터	6	7	8	9
실루엣 계수	0.654370813	0.655322172	0.654370813	0.637303729

표 2는 클러스터 개수 별 실루엣 계수를 나타낸다. 클러스터의 개수에 따른 각 실루엣 계수는 차이가 크지 않으며, 큰 의미를 가지고 있지 않다.

Ⅲ. 결론

본 논문은 K-means 클러스터링을 통해 랜섬웨어 주소 포함 비율을 보여준다. 클러스터 별 랜섬웨어 주소 포함 비율을 통해 랜섬웨어 주소와 뚜렷하게 구분되는 정상 클러스터를 알 수 있었다. 실험 결과로 109,807개의 주소는 정상적인 주소이며 더 많은 데이터가 있을 때 정답지로 활용할 수 있다. 또한 전체 주소의 약 10%가 랜섬웨어 주소로, 알려진 랜섬웨어 주소가 많지 않다. 그러나 클러스터의 61%를 차지한 랜섬웨어 주소가 아닌 39%의 주소도 알려지지 않은 랜섬웨어 주소라고 생각할 수 있다. 각 클러스터 별 실루엣 계수는 클러스터 개수 별 큰 차이가 없지만, 기존 주소 기준 데이터에서 추출한 특징을 현재보다 뚜렷한 차이를 갖는 특징을 추출하고 주소 데이터를 많이 갖고 있으면 실루엣 계수를 통해 클러스터가 몇 개일 때 클러스터링이 잘 되었는지 고려하고 판단할 수 있을 것으로 기대된다. 랜섬웨어 주소를 가진 클러스터가 뚜렷하게 구분되어 거래소에 제공한다면 비트코인을 환전할 수 없도록 해당 주소가 포함된 지갑을 차단하고 신원 확인 과정을 거치도록 사용할 수 있을 것이다.

참 고 문 헌

[1] LU, Tianliang, et al. Ransomware detection based on V-detector negative selection algorithm. In: 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). IEEE, 2017. p. 531-536.

[2] NAKAMOTO, Satoshi. Bitcoin: A peer-to-peer electronic cash system. Manubot, 2019.

[3] JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: a review. ACM computing surveys (CSUR), 1999, 31.3: 264-323.