KNOM Review '18-02 Vol.21 No.02
https://doi.org/10.22670//knom.2018.21.2.18

비트코인 네트워크 트랜잭션 이상 탐지를 위한 특징 선택 방법

백 의 준◆, 신 무 곤*, 지 세 현*, 박 지 태*, 김 명 섭°

The Method of Feature Selection for Anomaly Detection in Bitcoin Network Transaction

Ui-Jun Baek◆, Mu-Gon Shin*, Se-Hyun Jee*, Jee-Tae Park*, Myung-Sup Kim°

요 약

사토시 나타모토에 의해 블록체인 기술이 개발되고 비트코인이 새로운 암호화폐 시장을 개척한 이후 여러 암호화폐들이 등장하고 그 수와 규모는 나날이 증가하고 있다. 또한 블록체인 기술의 익명성과 여러 취약점을 이용한 범죄들이 발생하고 있으며 이에 취약점 개선과 범죄 예방을 위한 많은 연구들이 진행되고 있으나 범죄를 저지르는 사용자들을 탐지해내기엔 역부족이다. 따라서 네트워크 내 자금 세탁, 자금 탈취 등 이상 행위를 탐지 하는 것은 매우 중요하며 이에 본 논문에서는 비트코인 네트워크의 트랜잭션 및 유저 그래프의 특징들을 수집하고 이로부터 통계정보를 추출한 후 이를 로그 스케일 상에서 플롯으로 나타낸다. 시각화된 플롯을 Densification Power Law와 Power Degree Law에 따라 분석하고 결과적으로 비트코인 네트워크 내 비정상 트랜잭션 및 비정상 유저를 포함하는 이상 탐지에 적절한 특징들을 제시한다.

Key Words : Blockchian, Bitcoin, Feature Selection

ABSTRACT

Since the development of block-chain technology by Satoshi Nakamoto and Bitcoin pioneered a new cryptocurrency market, a number of scale of cryptocurrency have emerged. There are crimes taking place using the anonymity and vulnerabilities of block-chain technology, and many studies are underway to improve vulnerability and prevent crime. However, they are not enough to detect users who commit crimes. Therefore, it is very important to detect abnormal behavior such as money laundering and stealing cryptocurrency from the network.

In this paper, the characteristics of the transactions and user graphs in the Bitcoin network are collected and statistical information is extracted from them and presented as plots on the log scale. Finally, we analyze visualized plots according to the Densification Power Law and Power Law Degree, as a result, present features appropriate for detection of anomalies involving abnormal transactions and abnormal users in the Bitcoin network.

※이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2018-0-00539-001, 블록체인의 트랜잭션 모니터링 및 분석 기술개발)
※이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2018R1D1A1B07045742)
◆ First Author : Korea University Department of Computer and Information Science, pb1069@korea.ac.kr
° Corresponding Author : Korea University Department of Computer and Information Science, tmskim@korea.ac.kr
* Korea University of Department of Computer and Information Science, {tm0309, sxzer, pjj5846}@korea.ac.kr
논문번호 : KNOM2018-02-06, Received December 5, 2018; Revised December 16, 2018; Accepted December 21, 2018

Ⅰ. 서 론

사토시 나카모토에 의해 블록체인 기술이 개발되고 비트코인이 새로운 암호화폐 시장을 개척한 이후 여러 암호화폐들이 등장하였으며 그 수와 규모는 나날이 증가하고 있다. 2018년 12월 7일 기준으로 시장에서 거래되고 있는 암호화폐는 878종에 달하고 상위 10개 암호 화폐의 시가총액의 합은 110조원에 달한다. 이러한 암호화폐 시장의 급격한 성장에 따라 블록체인 기술의 취약점과 익명성을 이용하는 여러 범죄들이 발생하고 있으며 이러한 범죄들로 발생하는 피해는 막대하다. 취약점 개선과 범죄 예방 및 탐지를 위한 많은 연구들이 진행되고 있으나 악성 행위를 저지르는 악성 사용자들과 범죄들을 정확히 탐지해내기엔 역부족이다. 그러므로 블록체인 네트워크 내 발생하는 자금 세탁 및 암호화폐 탈취와 같은 악성 행위를 탐지하고 분석하는 것은 매우 중요한 연구 분야이다. 본 논문에서는 비트코인 네트워크 내 트랜잭션 데이터를 수집하고 트랜잭션과 유저 기준의 두 그래프를 추출하여 로그 스케일화 된 플롯으로 나타낸다. 이후 시각화된 플롯을 Densification Power Law 법칙에 따라 분석하고 비트코인 네트워크 내 발생하는 악성 행위를 분석하고 탐지하기에 적절한 특징들을 제시한다.

본 논문은 1장 서론에서 연구 배경과 연구 목표에 대해 서술하며 2장에서 관련 연구에 대해 서술한다. 3장 본론에서 핵심 개념 및 데이터 수집 및 통계를 추출하는 방법 및 추출된 데이터 구조, 플롯으로 나타내는 방법에 대해 서술하며 4장 실험 결과에서 수집 및 추출한 데이터를 플롯으로 나타내고 분석한 결과를 제시한다. 마지막으로, 5장에서 분석 결과를 설명하며 한계점과 향후연구에 대해 설명하고 본 논문을 마친다.

Ⅱ. 관련 연구

[1]은 시간에 따른 그래프 변화의 특성에 대해서 서술한다. [1]에서는 핵심 아이디어가 되는 Densification Power Law라는 법칙을 제시한다. 이는 실제 세상의 네트워크의 그래프의 에지 수와 노드 수가 로그 스케일 상에서 선형 함수의 형태를 띈다는 경험칙이다. 그리고 그래프의 In-Degree, Out-Degree등을 포함한 실제 네트워크의 많은 주요 특성을 재현하는 Foreset Fire 모델을 제시하는데 이는 주어진 그래프가 Densification Power Law와 그 특성을 만족하지 못한다면 그래프는 비정상적이라는 것이다. 또한 Power Law Degree 법칙을 제시하며 제시한 법칙에 따라 네트워크 내 악의적인 사용자들의 활동에 의해 발생한 분포를 발견한다면 이는 네트워크 내 비정상적인 활동이 일어난 것이라고 결론을 내릴 수 있다고 설명한다.

[2]에서는 [1]에서 제시한 법칙들을 이용하여 비트코인 네트워크 내 그래프를 분석하는 방법을 제안하였다. 또한 추출한 그래프를 K-means 알고리즘을 통해 클러스터링하고 클러스터 내 데이터 간 LOF(Local Outlier Factor) 수치를 계산하여 이상 징후를 가지는 유저 데이터와 트랜잭션 데이터를 추출하였다. 마지막으로 추출된 유저 데이터와 트랜잭션 데이터의 교차 검증하는 방법을 제안하였다.

[3]은 블록체인 네트워크를 분석하는 모듈형 프레임워크를 제안하며 이는 동일한 사용자 및 사용자 그룹에 속할 가능성이 높은 주소 및 사용자들을 분류하고 이러한 결과를 시각화 한다. 또한 제안한 방법을 통해 주소와 사용자 사이의 경로와 역-경로를 찾아 수동적인 조사를 지원하며 실제 사례 분석을 기반으로 불법적인 웹사이트인 SilkRoad의 지갑에 속할 가능성이 있는 주소를 식별하거나, CryptoLocker와 같은 랜섬웨어 피해자 및 피해 금액에 대한 정보를 정확히 수량화 하였다.

[4]는 블록체인 네트워크 데이터 수집부터 분석까지 일반 사용자에게 비트코인 네트워크의 정보를 제공하는 분석 플랫폼을 제안하였으며 트랜잭션 그래프를 클러스터링 알고리즘을 통해 네트워크 내 사용자와 주소 내 연관성을 추출하였다.

[5]는 데이터의 유용성을 높이기 위한 클러스터링과 엔트로피를 이용한 익명화 기법을 제안하였으며 서비스에서 활용될 데이터의 유용성을 높이기 위해 엔트로피를 이용한 K-anonymity를 제안하였다.

[3,4]는 모두 블록체인 네트워크를 분석하고자 하는 일반 사용자에게 Forensic 분석의 가능성을 제시할 순 있으나 Heuristic한 기준과 수동적인 분석으로 시시각각 변화하는 네트워크의 특성을 모두 반영하기 힘들다는 한계점을 지닌다. [2]는 비트코인 네트워크의 그래프 데이터 클러스터링과 클러스터 내 이상치인 LOF 계산을 통해 의심스러운 트랜잭션 혹은 비정상적인 사용자를 탐지하는 방법을 제안하였으나 탐지에 사용한 데이터가 비정상적인 사용자와 정상적인 사용자를 구분하기에 충분치 않은 일반적인 특징들로 이루어져있어 정확한 탐지가 어

렵다는 한계점을 지닌다. 따라서 본 논문에서는 비트코인 네트워크의 그래프의 클러스터링[3,4,5] 및 분석에 적절한 특징을 선택하기 위하여 비트코인 네트워크로부터 데이터를 수집하고 통계 데이터를 추출하여 이를 Densification Power Law와 Power Law Degree[1,2]에 따라 분석하고 향후 진행될 클러스터링에 적절한 특징들을 제시한다.

Ⅲ. 본 론

본 장에서는 핵심 개념과 데이터 수집 및 처리에 대한 전반적인 과정을 설명한다. 비트코인 네트워크로부터 트랜잭션 데이터를 포함한 데이터를 수집하고 이로부터 유저 데이터를 추출한다. 추출한 유저 데이터와 트랜잭션 데이터로부터 두 기준의 그래프를 추출하고 두 그래프의 통계 정보를 추출한다. 마지막으로, 두 그래프를 Densification Power Law와 Power Law Degree를 이용해 시각화하여 분석한다. 전체적인 개요는 그림 1과 같다.

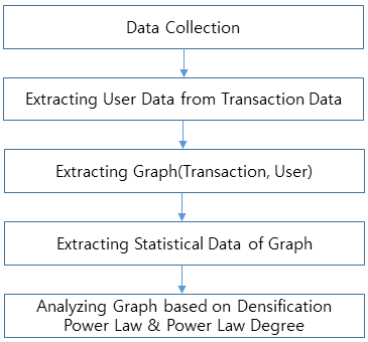

그림 1. 전체 실험 개요
Fig. 1. Outline of whole experiment

1. Power Law Degree & Densification Power Law

Densification Power Law는 노드 N과 에지 E로 이루어진 그래프에서 네트워크의 특정 시간의 t의 노드 개수의 a 제곱은 특정시간 t의 에지의 개수에 비례한다는 법칙이며 이는 수식 1과 같다.

$$E(t) \propto N(t)^{a} \quad (1)$$

Power Law Degree는 실제 정상적인 네트워크에서 P(K)를 차수 k를 가지는 노드의 특징이라고 정의하고 c가 양의 정수일 때 P(k)는 차수 k의 역수에 비례한다는 법칙이며 이는 수식 2와 같다.

$$P(k) \propto k^{-\gamma} \quad (2)$$

P(k)는 유저의 잔액, 평균 트랜잭션 사이즈, 트랜잭션의 총 거래금액 등 노드의 특징으로 대체될 수 있으며 본 논문에서는 기존 연구에서 사용했던 특징들의 일반적인 정보뿐만 아니라 총 합, 최댓값, 최솟값, 평균, 표준편차 등 통계정보를 추출하여 이러한 특징들을 사용한다.

2. 데이터 수집

비트코인 네트워크의 첫 번째 블록부터 200,000 높이의 블록에 담긴 트랜잭션 데이터를 수집하였으며 수집한 트랜잭션 데이터를 재-정렬하여 유저 데이터를 추출하였다. 추출한 트랜잭션 데이터와 유저 데이터는 그림 2와 같다.

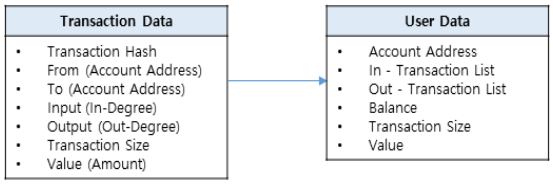

그림 2. 추출한 트랜잭션 데이터와 유저 데이터
Fig. 2. Extracted transaction and user data

수집한 트랜잭션 데이터는 트랜잭션의 해쉬, 입력을 받은 유저의 정보인 Input(In-Degree), Output(Out-Degree), 트랜잭션의 사이즈 그리고 거래금액인 Value로 이루어져 있다. 트랜잭션 데이터로부터 추출한 유저 데이터는 유저의 식별자인 주소, 입력 받은 트랜잭션의 리스트, 출력한 트랜잭션의 리스트, 잔액, 유저가 포함된 트랜잭션의 사이즈, 입력 값과 출력 값의 합인 Value로 구성하였다.

수집하고 추출한 트랜잭션 데이터와 유저 데이터로부터 트랜잭션 기준의 유저-트랜잭션 그래프와 유저 기준의 유저-트랜잭션 그래프를 추출하였다. 비트코인 네트워크의 트랜잭션은 한 개 이상의 입력과 출력을 가질 수 있으므로 트랜잭션 기준의 그래프와 유저 기준의 그래프는 모두 방향성을 가지는 차수(In-Degree, Out-Degree)를 가진다. 추출한 두 기준의 그래프는 그림 3, 4와 같다.

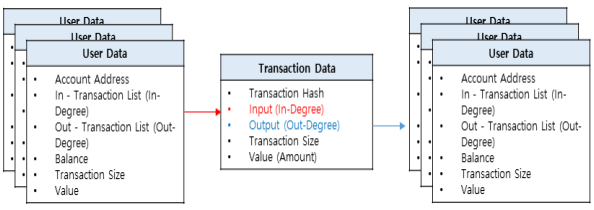

그림 3. 추출한 트랜잭션 기준의 그래프
Fig. 3. Extracted transaction graph

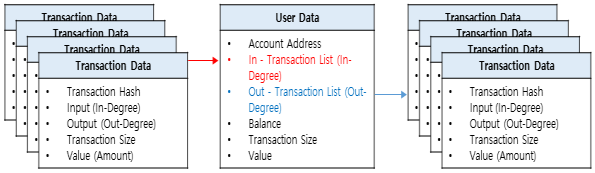

그림 4. 추출한 유저 기준의 그래프
Fig. 4. Extracted User Graph

3. 수집 데이터 통계 추출

추출한 트랜잭션 및 유저 그래프의 특징들로부터 통계 데이터를 추출한다. 추출한 모든 통계 데이터는 표 1,2와 같다.

표 1. 추출한 유저 그래프의 통계 정보
Table 1. Extracted statistical of user graph

User Graph User-Transaction	In-Degree	Value	Sum
			Max
			Min
			Mean
			Stdvar
		Transaction Size	Sum
			Max
			Min
			Mean
			Stdvar
	Out-Degree	Value	Sum
			Max
			Min
			Mean
			Stdvar
		Transaction Size	Sum
			Max
			Min
			Mean
			Stdvar

표 2. 추출한 트랜잭션 그래프의 통계 정보
Table 2. Extracted statistical of transaction graph

Transaction Graph Transaction - User	In-Degree	Amount	Sum
			Max
			Min
			Mean
			Stdvar
	Out-Degree	Amount	Sum
			Max
			Min
			Mean
			Stdvar

유저 기준의 그래프에서는 그래프를 이루는 각 In-Degree, Out-Degree의 트랜잭션 거래량과 트랜잭션 사이즈로부터 5개의 통계 정보를 추출한다. 이 과정을 통해 총 20가지의 유저그래프 통계 데이터가 추출된다.

트랜잭션 기준의 그래프에서는 그래프를 이루는 각 In-Degree, Out-Degree의 트랜잭션 거래량으로부터 5개의 통계 정보를 추출한다. 이 과정을 통해 총 10가지의 트랜잭션 그래프 통계 데이터가 추출된다.

4. 데이터 시각화

수집하고 추출한 두 기준의 그래프를 비교가 용이하도록 로그 스케일의 플롯으로 나타낸다. 모든 플롯의 y 축은 공통적으로 차수(Number of In-Degree, Number of Out-Degree)로 설정하였고 x 축은 추출한 In-Degree, Out-Degree 특징들의 통계 정보로 설정하였다.

Ⅳ. 실험 결과

본 장에서는 추출한 데이터를 그래프로 나타내고 이에 대해 설명한다. 서론에서 언급했듯이 그래프의 분포가 비선형일 경우 해당 네트워크 내 이상이 있다고 판단할 수 있다. 따라서 그래프의 분포를 보며 분석하고 비선형적인 그래프의 분포를 가지는 특징들을 찾는다.

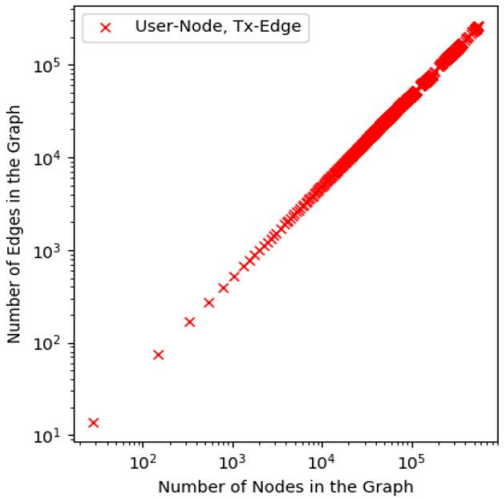

그림 5. 노드 수와 에지 수 플롯 : 유저 그래프
Fig. 5. Number of node - Number of Edge Plot : User graph

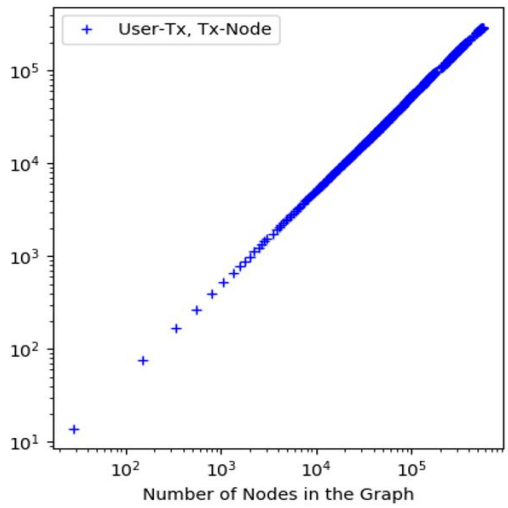

그림 6. 노드 수와 에지 수 플롯 : 트랜잭션 그래프
Fig. 6. Number of node - Number of edge plot : Transaction graph

그림 5, 6은 두 기준의 그래프의 노드 수와 에지 수를 로그 스케일 상에서 나타낸 것이며 두 그래프 모두 선형함수의 형태를 띠는 것을 확인할 수 있다. Densification Power Law에 따라 분석하면 이를 정상적인 네트워크라고 판단할 수 있으나 비트코인 네트워크 내에는 분명한 비정상적인 행위가 발생하고 있으므로 이는 노드 수와 에지의 수의 정보로는 비정상적인 행위 혹은 주체를 탐지하기 힘들다고 판단할 수도 있다.

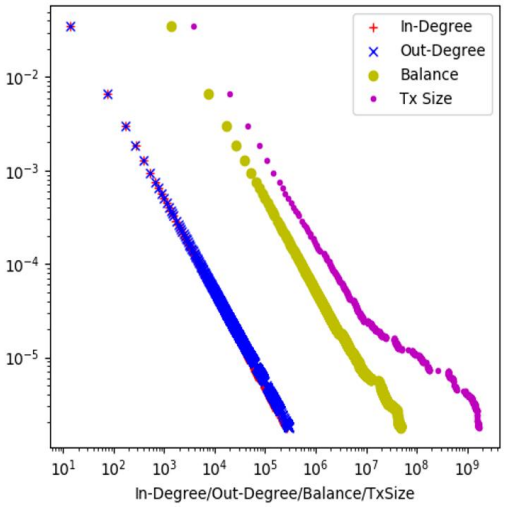

그림 7. In/Out-Degree 특징 : 유저 그래프
Fig. 7. In/Out-Degree feature : User graph

그림 7, 8은 두 기준의 그래프의 특징 중 In-Degree와 Out-Degree의 수, 잔액, 트랜잭션 사이즈, 총 거래금액 등 일반적인 정보를 이용하여

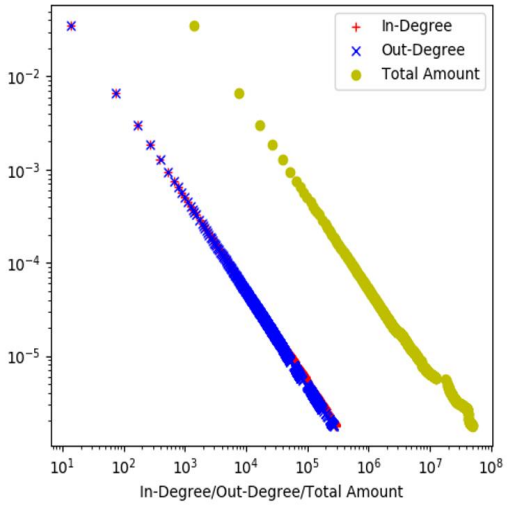

그림 8. In/Out-Degree 특징 : 트랜잭션 그래프
Fig. 8. In/Out-Degree feature : Transaction graph

나타낸 것이다. 이를 Power Law Degree에 따라 분석하면 In-Degree 특징 플롯에서 트랜잭션의 분포를 제외하고 모든 그래프가 선형함수의 형태를 띠는 것을 확인했으며 이러한 일반적인 정보를 통해서는 정상과 비정상을 구분할 명백한 특징이 부족하다고 말할 수 있으며 정확한 비정상 행위 및 주체 탐지가 어렵다고 판단할 수 있다.

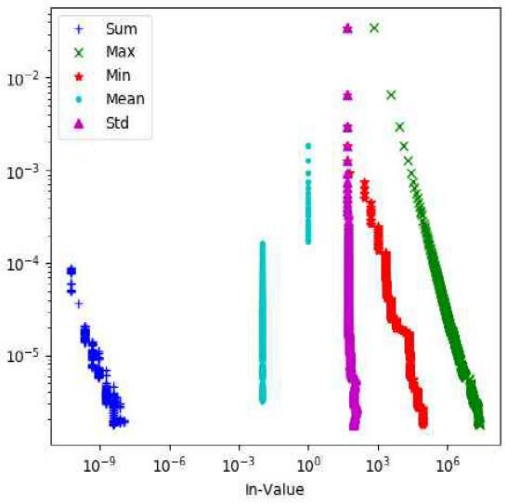

그림 9. In-Degree value 통계정보 : 트랜잭션 그래프
Fig. 9. Statistical data of In-Degree Value : Transaction graph

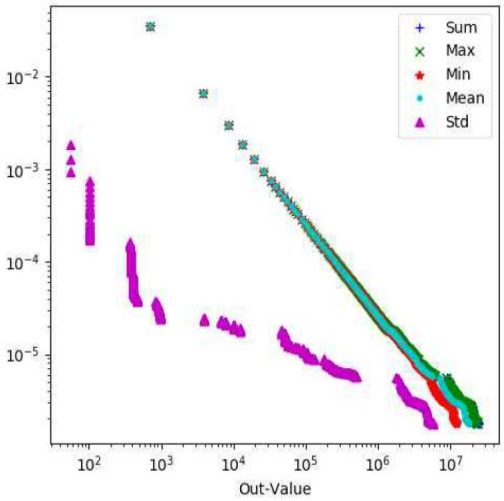

그림 10. Out-Degree value 통계정보 : 트랜잭션 그래프
Fig. 10. Statistical data of In-Degree Value : Transaction graph

그림 9, 10은 트랜잭션 기준 그래프의 특징 중 In-Degree와 Out-Degree의 Value 값의 시간에 따른 분포를 나타낸 것이며 In-Degree의 평균과 Out-Degree의 표준편차에서 비선형적인 형태를 띠는 것을 확인하였다.

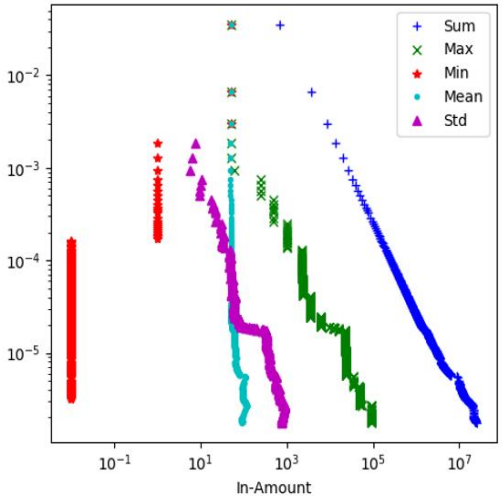

그림 11. In-Degree value 통계정보 : 유저 그래프
Fig. 11. Statistical data of In-Degree Value : User graph

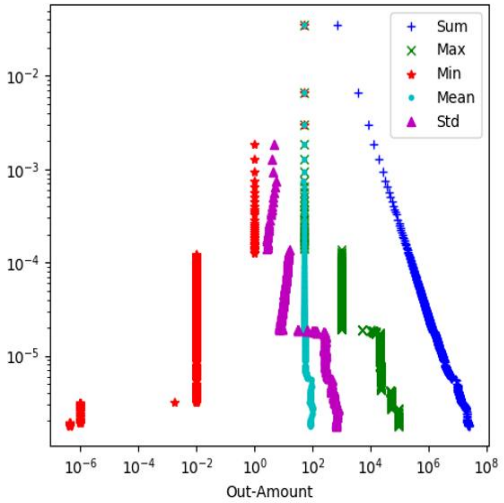

그림 12. Out-Degree value 통계정보 : 유저 그래프
Fig. 12. Statistical data of Out-Degree Value : User graph

그림 11, 12는 유저 기준 그래프의 특징 중 In-Degree와 Out-Degree의 Value 통계 정보의 시간에 따른 분포를 나타낸 것이며 In-Degree의 최솟값, 최댓값, 표준편차와 Out-Degree의 최솟값, 최댓값, 표준편차에서 비선형적인 형태를 띠는 것을 확인하였다.

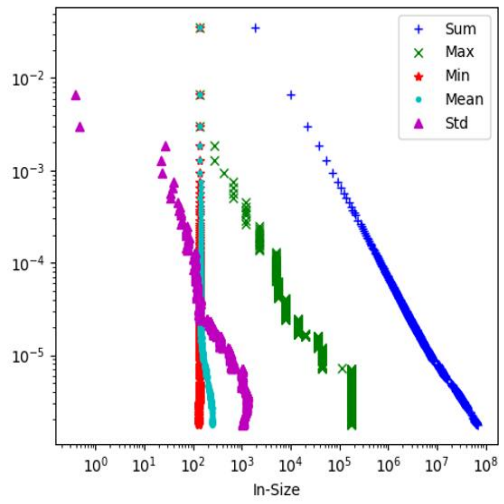

그림 13. In-Degree 트랜잭션 사이즈 통계정보 : 유저 그래프
Fig. 13. Statistical data of In-Degree transaction size : Transaction graph

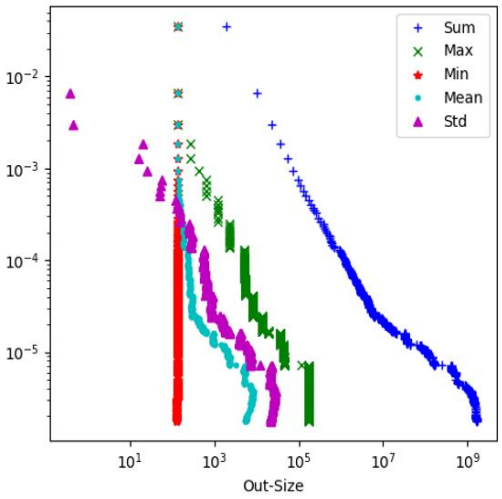

그림 14. Out-Degree 트랜잭션 사이즈 통계정보 : 유저 그래프
Fig. 14. Statistical data of Out-Degree transaction size : User graph

그림 13, 14는 유저 기준 그래프의 특징 중 In-Degree와 Out-Degree의 트랜잭션 사이즈 통계정보의 시간에 따른 분포를 나타낸 것이며 In-Degree의 최댓값, 평균, 표준편차와 Out-Degree의 총 합, 최댓값, 평균, 표준편차에서 비선형적인 형태를 띄는 것을 확인하였다.

그림 [9-14]는 유저 및 트랜잭션 그래프의 특징들로부터 통계정보를 추출하여 이를 로그 스케일 상의 플롯으로 나타낸 것이며 이로부터 그림 [5-8]의 선형적인 형태와는 비선형적인 구간을 가지는 것을 확인하였다. Densification Power Law와 Power Law Degree를 통해 분석하였을 때, 이는 네트워크에 비정상적인 행위가 있을 수 있다고 해석할 수 있다. 이러한 그래프의 통계정보를 특징으로 사용하였을 경우 비선형적인 구간을 실제사례를 바탕으로 분석할 때 유저 혹은 트랜잭션이 정상인지 비정상인지 구분하기에 적절한 분포를 가진다고 판단된다.

V. 결 론

본 논문은 비트코인 네트워크의 트랜잭션 데이터와 유저 데이터를 수집 및 추출하고 이로부터 두 기준의 그래프를 추출하였다. 또한 추출한 두 기준의 그래프의 특징들로부터 통계 정보를 추출하였고 이를 플롯으로 시각화하여 분석하였다. 분석 결과를 통해 수집 가능한 일반적인 정보보다 통계정보가 명확히 구분할 수 있는 분포를 띈다는 것을 확인하였다. 이러한 명확히 구별되는 특징을 가진 그래프들을 클러스터링 알고리즘을 통해 분류하고 실제 비트코인 네트워크 내에서 발생했던 실제 사례들과 비교 분석을 통해 비정상적인 트랜잭션과 유저가 어떠한 특성을 지니고 있는지 분석할 수 있었다. 이를 통해 기계학습 기반의 K-means와 심층학습 기반의 SOM(Self-Organized Map)과 같은 클러스터링 알고리즘을 이용하여 네트워크 내 비정상적인 행위와 행위의 주체인 악의적인 사용자들을 탐지하는 이상 탐지의 가능성을 제시하였다. 하지만 그래프의 분포가 비선형적인 것을 확인하였음에도 실제 사례 기반의 비교 및 상세 분석이 수행되지 않았기에 비선형적인 구간에서 이상이 있다고 판단하기 어렵다는 명백한 한계점을 지닌다. 이러한 한계점을 극복하기 위하여 자금 탈취 혹은 세탁과 같은 비정상적인 행위에 대한 실제 사례를 수집하고 이와 비교 분석하여 본 논문에서 분석하고 제시한 통계 데이터의 가용성을 검증할 예정이며 분석할 수 있는 특징들에 대해 추가적으로 분석할 예정이다.

References

[1] Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. "Graph evolution: Densification and shrinking diameters." ACM Transactions on Knowledge Discovery from Data (TKDD) 1.1 (2007): 2.

[2] Pham, Thai, and Steven Lee. "Anomaly Detection in the Bitcoin System-A Network Perspective." arXiv preprint arXiv:1611.03942 (2016)

[3] Spagnuolo, Michele, Federico Maggi, and Stefano Zanero. "Bitiodine: Extracting intelligence from the bitcoin network." International Conference on Financial Cryptography and Data Security. Springer, Berlin, Heidelberg, 2014.

[4] Kalodner, Harry, et al. "BlockSci: Design and applications of a blockchain analysis platform." arXiv preprint arXiv:1709.02489 (2017).

[5] Da Eun Lee, Choong Seon Hong, "A Study on Anonymity Scheme Using Entropy and Clustering", KNOM Review, Vol. 19, No. 1, pp. 22-30., Aug. 2016

백 의 준 (Ui-Jun Baek)

2018 고려대학교 컴퓨터정보학과 학사
2018년 - 현재 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

신 무 곤 (Mu-Gon Shin)

2012년 - 현재 고려대학교 컴퓨터정보학과 학사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

지 세 현 (Se-Hyun Jee)

2018 고려대학교 컴퓨터정보학과 학사
2018년 - 현재 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

박 지 태 (Jee-Tae Park)

2017 고려대학교 컴퓨터정보학과 학사
2017년 - 현재 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

김 명 섭 (Myung-Sup Kim)

1998 포항공과대학교 전자계산학과 학사
2000 포항공과대학교 전자계산학과 석사
2004 포항공과대학교 전자계산학과 박사
2006 Dept. of ECS, Univ of Toronto Canada
2006 - 현재 고려대학교 컴퓨터정보학과 교수
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크