A First Estimation of the Proportion of Cybercriminal Entities in the Bitcoin Ecosystem Using Supervised Machine Learning

Submitted by grigby1 on Mon, 03/05/2018 - 1:14pm

Title	A First Estimation of the Proportion of Cybercriminal Entities in the Bitcoin Ecosystem Using Supervised Machine Learning
Publication Type	Conference Paper
Year of Publication	2017
Authors	Yin, H. Sun, Vatrapu, R.
Conference Name	2017 IEEE International Conference on Big Data (Big Data)
Date Published	dec
ISBN Number	978-1-5386-2715-0
Keywords	Bagging classifiers, bitcoin, Bitcoin ecosystem, Bitcoin transactions, blockchain, composability, Computer crime, cryptocurrency, cyber-criminal entities, Cybercrime, cybercrime-related categories, cybercriminal entities, Ecosystem, Ecosystems, financial data processing, Gradient Boosting classifiers, gradient methods, illegal goods trading, illicit activities, learning (artificial intelligence), machine learning, Malware, Metrics, pattern classification, Peer-to-peer computing, peer-to-peer payment system, proportion estimation, pubcrawl, Public key, ransomware, resilience, Resiliency, security of data, supervised learning, supervised learning classifiers, supervised machine learning, uncategorised observations

Abstract

Bitcoin, a peer-to-peer payment system and digital currency, is often involved in illicit activities such as scamming, ransomware attacks, illegal goods trading, and thievery. At the time of writing, the Bitcoin ecosystem has not yet been mapped and as such there is no estimate of the share of illicit activities. This paper provides the first estimation of the portion of cyber-criminal entities in the Bitcoin ecosystem. Our dataset consists of 854 observations categorised into 12 classes (out of which 5 are cybercrime-related) and a total of 100,000 uncategorised observations. The dataset was obtained from the data provider who applied three types of clustering of Bitcoin transactions to categorise entities: co-spend, intelligence-based, and behaviour-based. Thirteen supervised learning classifiers were then tested, of which four prevailed with a cross-validation accuracy of 77.38%, 76.47%, 78.46%, 80.76% respectively. From the top four classifiers, Bagging and Gradient Boosting classifiers were selected based on their weighted average and per class precision on the cybercrime-related categories. Both models were used to classify 100,000 uncategorised entities, showing that the share of cybercrime-related is 29.81% according to Bagging, and 10.95% according to Gradient Boosting with number of entities as the metric. With regard to the number of addresses and current coins held by this type of entities, the results are: 5.79% and 10.02% according to Bagging; and 3.16% and 1.45% according to Gradient Boosting.

URL http://ieeexplore.ieee.org/document/8258365/
DOI 10.1109/BigData.2017.8258365
Citation Key yin_first_2017

Bagging classifiers bitcoin Bitcoin ecosystem Bitcoin transactions blockchain composability Computer crime cryptocurrency cyber-criminal entities Cybercrime cybercrime-related categories cybercriminal entities Ecosystem Ecosystems financial data processing Gradient Boosting classifiers gradient methods illegal goods trading illicit activities learning (artificial intelligence) machine learning malware Metrics pattern classification Peer-to-peer computing peer-to-peer payment system proportion estimation pubcrawl Public key Ransomware resilience Resiliency security of data supervised learning supervised learning classifiers supervised machine learning uncategorised observations