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

Haohua Sun Yin[1], Ravi Vatrapu[1,2]
[1]*Centre for Business Data Analytics, Copenhagen Business School, Denmark*
[2]*Westerdals Oslo School of Arts, Comm & Tech, Norway*
awasunyin@gmail.com, vatrapu@cbs.dk

***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.**

***Keywords*-Bitcoin; Blockchain, Cryptocurrency, Ecosystem, Cybercrime, Machine Learning, Supervised Learning, Ransomware**

I. Introduction

Bitcoin is a peer-to-peer payment system and digital currency, conceived in 2008 [1]. In 2015, Bitcoin was estimated to be accepted as a payment method by over 100,000 merchants around the world [2]. As a recent study showed, Bitcoins popularity, together with other digital currencies, continues to raise: there are 2.9 to 5.8 million unique users, majority of which are Bitcoin users [3]. Nonetheless, a significant slice of the Bitcoin ecosystem is, and has been, associated with illicit activities such as money laundering, cyber-extortion, thievery, scamming, terror financing or illegal goods trading in the clear through the darknet [5].

The name of Bitcoin has frequently shared headlines with darknet markets, malware (ransomware), and fraudulent acts in many media sources. Notorious examples of the above are respectively: the Silk Road 1.0 shut-down in 2013 and the rumoured exit scam of *AlphaBay* in July 2017; *WannaCry 2.0*s worldwide spread in May 2017, which impacted large organisations such as Telefnica or Deutsche Bahn; or *NotPetya* in June 2017, which affected more than 80 companies in Ukraine; and the famous cases of *Mt. Gox*s stolen Bitcoins worth $350 million in 2014 or the case of *Bitfinex*, another exchange that was hacked and lost around $60 million in 2016.

Due to Bitcoins characteristics, especially its pseudo-anonymity, the fact that it became the preferred payment system for illicit activities comes as no surprise. In contrast to other digital payment methods such as debit or credit card payments, Bitcoin transactions are not linked to real-world identities, but only to public keys or addresses, and the generation of the latter does not require any verified personal information. Whereas some entities voluntarily reveal their addresses when it is necessary to provide their services (e.g. donation addresses), others carry out privacy-enhancing payment schemes or leverage mixing services to obscure their spending habits. This behaviour is commonly seen in entities that are related to tor markets, ransom payments, scams, and thievery.

Regardless of anyone being able to generate Bitcoin addresses without linking any identity information and the possibility of obscuring their spending habits through privacy overlays or mixing services, the most common way of converting Bitcoin into fiat currencies is via exchange services such as *Kraken* or *Coinbase*, where customers must provide personal information in order to create an account and access to their services.

Since 2013, Bitcoins adoption skyrocketed and is estimated to grow steadily. This, combined with the fact that a portion of the ecosystem is involved with criminal activities, has attracted the attention of regulatory and law enforcement bodies as well as businesses. Businesses that interact directly or indirectly with Bitcoin seek for tools that will contribute to their compliance with local regulations. US companies, for instance, must comply with Anti Money Laundering (AML) and Know Your Customers (KYC) regulations, meaning that for transactions that involve certain amounts of Bitcoins valued in USD, the company must include the risk profile assessment of their customers in their compliance reports as required by local regulations and

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

authorities. Regulators seek access to data-driven sources that will provide them with a better view of the ecosystem, and law enforcement bodies seek tools that complement their analysis and contribute to their investigations on cybercrime-related cases as well as a better prevention framework.

The main motivation of the current research is to increase the transparency of the ecosystem, encouraging businesses and consumers to adopt Bitcoin as a payment system, growing its economy without the need of illicit activities. Furthermore, the results of the current research can benefit regulators, businesses, and law enforcement as follows: regulators who seek data-driven sources that estimate the Bitcoin landscape and its evolution; businesses that seek tools that contribute to their compliance and better risk assessment of Bitcoin transactions; and law enforcement bodies which seek tools that contribute to their analysis, investigations (linking Bitcoin addresses to illicit entities), and prevention (flagging current or future addresses that are likely to be associated with illicit activities).

Unlike prior publications (see section II), the current research utilises data that has already been enriched with different clustering techniques and heuristics provided by *Chainalysis* (data provider). This means that labelled examples of darknet markets, ransom payments, scams, and thievery are part of the training set. In this paper, we builds on [4] that showed that supervised learning techniques were appropriate for classification of the Bitcoin entities. With regard to practical applications, we limit the current scope of this paper to the rapid prototyping of a tool, which relies on a supervised learning model, that can contribute to cybercrime investigations and prevention.

Considering all of the above, this paper aims to answer the following research question and sub-questions: *How large is the share of cybercrime-related entities in the Bitcoin ecosystem? How many entities and addresses are related to fraudulent activities? How many Bitcoin are currently held by these entities?*

II. Related Work

The most relevant related work can be grouped in two: first, cybersecurity and cybercrime analysis; and second, articles aiming to de-anonymise the Bitcoin blockchain via data analysis.

A. Cybersecurity & Cybercrime Analysis

In this group, we find published reports by cybersecurity firms, that issued many reports focussing on analysing malware such as *Trojan.Ransomlock* before Bitcoins appearance, *Cerber*, *CryptXXX*, and *Locky* ransomware families after Bitcoins adoption as ransom payment system. Afterwards, the reports' scope expanded to include security breaches, phishing, targeted attacks, in addition to ransomwares among other malware. [6][7]

Within the same group, there are academic publications that aim to analyse malware or specific services in the Bitcoin ecosystem. Notable examples of the first are: the in-depth analysis of notorious families of ransomwares such as *CryptoWall* and *CryptoLocker*[8]; or the usage of open sourced data from BitcoinTalk and Reddit, clustering a total of 968 Bitcoin addresses and identifying 795 ransom payments worth 1,128.40 BTC ($310,472.38)[9]. Notable examples of the second are: the analysis of the infamous darknet market named The Silk Road[10]; the analysis of mixing services (Bitcoin Fog, BitLaundry, and the Send Shared feature from Blockchain.info), which obfuscate the source of Bitcoin transactions for their customers, through transaction graph analysis[11].

B. Articles Aiming to De-Anonymise the Bitcoin Blockchain

There have been numerous publications that aimed to challenge Bitcoins assumed pseudo-anonymity. The first approach to unraveling Bitcoins anonymity applied network analysis techniques on addresses crossed with open source information sourcing from Wikileaks, for instance, revealing that it is possible to connect Bitcoin user addresses with each other[12]. A second approach involved the direct interaction with the network by sending transactions and by clustering public keys following co-spend heuristics, which concluded with the identification of 1.9 million Bitcoin addresses connected to real services or pseudo-identities (nicknames)[13]. An open-source framework was designed to parse the Bitcoin Blockchain, cluster public keys, label the clusters and visualise the network. The model was tested and resulted in the identification of an address containing 111,114 BTC belonging to a Silk Road cold wallet and the accurate quantification of ransoms paid to CryptoLocker with only an address posted by a victim on a forum as a lead[14]. Another approach was to apply statistical analysis in order to identify its users behavioural patterns when sending, receiving or storing coins which found that the vast majority of coins remain stored in addresses that have never been involved in outgoing transactions, in contrast with large amounts of transactions moving small amounts of coins and the particular subject of analysis, hundreds of transactions that send more than 50,000 BTC at once[15]. Finally, unsupervised learning methods (K-means algorithm) were applied to cluster a portion of the Bitcoin Blockchain aiming to detect anomalous behaviour from mixing services transaction data, uncovering anomalous transactions, as well as identifying unusual activity from some users, suspected to be laundering money[16].

III. Conceptual Framework

A. Introduction to Blockchain and Bitcoin

A blockchain is a constantly growing list of blocks which are the format a collection of record (transactions) is stored. Blocks contain a hash, a link to the previous block, a timestamp, and transaction data, secured through cryptography. It is distributed as it is run and controlled by many peers or nodes, and decentralised, as there is no central authority controlling it. In public blockchains such as Bitcoin and Ethereum, any peer can become a validator or a miner, and anyone can download and store a copy of its entire history. The term broadcasted refers to the fact that once a block has been mined, the information of this event is sent to all the nodes of the network and that the transaction(s) has been confirmed, irreversible, visible for the public.[17]. Bitcoin is a digital currency, often compared to cash due to its pseudo-anonymity as its users identity are hidden behind a pseudonyms instead of real world identities, such as first and last names. Bitcoin relies on blockchain technology, which is by design resistant to data corruption and enables peer-to-peer transactions that are verified by each of the nodes on the network[1].

1) Public keys, private keys, and wallets: Many elements are involved when making a transaction using Bitcoin. Starting with the address (public key or pseudonym), generation of which relies on two cryptographic primitives: ECDSA (digital signature scheme) and SHA-256 (one-way hash function). For every public key, a pairing private key is generated which is necessary to sign transactions. Every user can easily generate as many key pairs as desired, but due to the nature of these keys, wallets are often used to manage them. Bitcoin wallets are software (hot wallets) or combination of software and hardware (cold wallets) that allows users to generate, store and use keys in a convenient way.

2) Bitcoin transactions: In a simple scenario where one party wishes to send BTC to another party, the sender needs to create a request where the amount and the receiver address are specified. Afterwards, the private key together with the public key is used to create a signature for the request, and the transaction is shared with the peers on the network. After receiving it, each peer verifies the signature and that the amount of coins have not been previously spent (avoiding double spending is crucial for the system). Once the transaction has passed these two verifications, they are aggregated into blocks, providing a timestamp for every transaction that it includes. The word mining refers to the process of creating these blocks. When miners, users or parties who have economic incentives to create blocks, successfully create a block, the transaction is finally broadcasted to the network and, hence, is public. It is important to note about change addresses: in Bitcoin, whenever a certain amount is sent to another address, in reality all the coins of that address are sent out. However, the part that was intended to be sent to another party is not the only amount in circulation but also the remaining coins are either sent back to the same address or to a newly-generated one, depending on the wallet (some programs create a new address automatically) or user preferences.

B. Clusters, entities, and categories

A cluster is a group of addresses, ownership of which belongs to one entity. Ownership is determined by the access to the addresses corresponding private keys, and hence, the control over the funds that are stored in them. An entity is a person or organisation that exists in the real world. Categories refer to types of entities, which are defined depending on the main activity of the entity. In our current study, the categories available in the dataset are not only tor markets, scams, ransomware, mixing, and stolen bitcoins, but also exchange, gambling, merchant services, hosted wallets, mining pools, personal wallets, and other.

1) Categories' description and examples:

- Tor Market: Black markets primarily facilitating trading of legal or illegal goods like narcotics, stolen credit cards, passports, etc. These sites are only accessible on the deep web through e.g. the TOR-browser. Examples: The Silk Road, Alphabay.
- Scam: Entities that deceive their customers by impersonating an existing service or pretending to provide a service in order to steal their Bitcoin. Examples: Bitdaytrade, Bitcoin7.
- Ransomware: Entities that are utilising the Bitcoin Blockchain as a payment system to receive ransom fees. Examples: WannaCry 2.0, NotPetya.
- Mixing: Entities that apply techniques to reduce the traceability of their clients transactions as a service. Examples: Bitcoin Fog, BitMixer.
- Stolen Bitcoins: Entities that managed to gain access to the private key(s) owned by other entities and committed thievery. Examples: Bitcoinica, BTC-Es hack.
- Exchange: Entities that allow their customers to trade fiat currencies for Bitcoins and vice-versa. Examples: Coinbase, Kraken.
- Gambling: Entities that offer gambling services that accept Bitcoin. Examples: Lucky Games, Nitrogen Sports.
- Merchant Services: Entities that offer solutions to businesses in order to facilitate the adoption of Bitcoins as a payment method for their customers. Examples: Purse.io, BitPay.
- Hosted-Wallet: Trusted entities that offer Bitcoin storage as a service. Examples: Xapo, Bitcoin Wallet.
- Mining Pool: Entities composed by distributed miners who share their processing power over a mining network and gain a compensation that equals to their contribution in solving a block. Examples: AntPool, BTC Top.
- Personal Wallet: Addresses or group of addresses managed by one entity for private uses such as trading, buying goods, gambling, etc.
- Other: Entities that have been identified but do not belong to any of the categories mentioned above as

they provide different services. Examples: WikiLeaks donation address or Secure VPN.

C. *Data provider's clustering methodology*

The Bitcoin transaction data is publicly accessible either directly downloading the entire blockchain or using free block explorers. Nonetheless, the observations in the dataset are not individual transactions but clusters. The data provider has clustered, identified, and labelled addresses through the following means: Co-spend clustering, whenever two or more input addresses are used for one transaction; intelligence-based clustering, where information and intelligence outside of the blockchain obtained via data partnerships is used; and behavioural clustering, where clustering is done according to known patterns that are dictated by the wallet software or systems used.

IV. Methodology

A. *Dataset description*

The data can be split in two subsets: the categorised dataset, where the label for each observation is one of the twelve mentioned above, contains a total of 874 observations; and the uncategorised dataset, which have the label of uncategorised as a placeholder, contains. In both cases, the following data for every cluster is available:

- Transactions: information about the entire transaction history of the cluster, some of its columns are: transaction hash, timestamp, input address, output address or value.
- Addresses: a collection of all addresses that have been grouped into this cluster, hence belonging to one entity. The columns are address, number of transactions with each peer address, and value.
- Counterparties: includes the history of parties that interacted directly with the specific cluster, meaning that the current cluster has sent or received directly from these addresses at least once. The columns are the counterparty address, the value and the category, if available, of the counterparty.
- Exposure: provides percentages that represent how much direct input or output the current cluster has with a certain type of entity. For example, if 30 of 100 sending transactions of cluster X go to an exchange, then one of the rows of this dataset will say 30% direct sending exposure to the category exchange.

To summarise the dataset, Figure 1 shows the category distribution, where the *personal wallet* and *exchange* categories are more abundant than the rest. Table I present statistics of the datasets by categories, aiming to provide an overview of how many transactions and addresses each cluster includes.

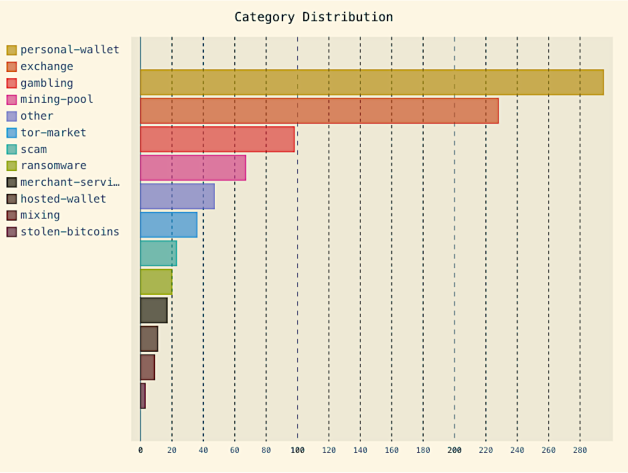

Figure 1. Category Distribution of the Dataset

B. *Data preprocessing*

1) Data cleaning: The data cleaning phase refers to the tasks performed on the datasets described above. Since these have already been enriched by the Data Provider and designed to be readable by any customer, there was very little cleaning needed. As Machine Learning algorithms normally do not handle null (NaN), infinity (inf) and dangerous operations such as dividing by zero, the main task was to remove empty cells from the datasets. This has been done as follows:

- If the value is normally an integer, it will be replaced by 0. For example: In the Transactions dataset, there are two columns, one for the output value of the transaction in BTC and another one for the input value in BTC. When the transaction is only outgoing, cluster X sends BTC to cluster Y, the input value in BTC cell is null, which is at this stage replaced by a 0.
- If the value is normally a float, it will be replaced by 0.0. For example: In the Exposure dataset, whenever a Cluster X has no interaction at all with scams, there will be a blank in the corresponding row, which will be replaced by 0.0.
- If the value is normally a string, it will be replaced by the most appropriate string depending on the column. Even though this might seem a risky practice, it only happens in this scenario: in the Counterparties dataset, if cluster X interacts with cluster Y, and the latter has not been categorised, there will be a blank in the cell, which will be replaced by uncategorised.

2) Dimensionality reduction: The two techniques used to generate the attributes from the dataset are manual feature extraction and feature engineering. A total of 99 attributes are generated, of which Table II provides an overview.

Table I
SUMMARY STATISTICS OF DATASETS

Categories	Observations	Mean TX	Median TX	Min. TX	Max. TX	Mean Address	Max. Address
scam	23	6399852	39853	5	31966333	2624782	29216558
mixing	9	2573113	10422	66	8056071	56149	222107
merchant-service	17	2002106	5638	32	13409499	133787	1190042
hosted-wallet	11	1571594	124297	69	5410610	901522	4596261
mining-pool	67	415715	163050	77	25694891	180435	10401844
tor-market	36	338927	37835	137	5507480	177097	2459741
gambling	98	143257	65442	17687	13075255	14069	1079523
exchange	228	137152	18250	58	12968079	19010	670420
ransomware	2	24411	19521	960	68050	528	7902
stolen-bitcoins	3	1260	1041	118	2622	699	2072
other	47	931	407	11	14134	581	8526
personal-wallet	295	177	38	3	5419	71	1197
uncategorised	100000	9724	1	1	31974832	3312	29235986

Table II
SAMPLE OF ATTRIBUTES

Feature Name	**Process**
TRX_btc_rcvd_sum	Extraction
TRX_btc_sent_sum	Extraction
TRX_balance	Extraction
TRX_usd_rcvd_sum	Extraction
TRX_withdrawals	Extraction
TRX_deposits	Extraction
TRX_clusterLT	Engineering
TRX_btc_rcvd_median	Extraction
TRX_sent_exp_scam	Engineering
TRX_sent_exp_ransomware	Engineering
TRX_rcvd_exp_mixing	Engineering
TRX_rcvd_exp_gambling	Engineering
TRX_rcvd_exp_tor_market	Engineering
TRX_rcvd_exp_uncategorised	Engineering
ADD_cluster_addresses	Extraction
CP_unique_counterparties	Extraction
CP_trx_count_sum	Extraction
CP_btc_sent_mean_perPeer	Engineering
CP_trx_count_perPeer	Engineering
CP_btc_flow_perPeer	Engineering

Table III
PERFORMANCE OF 13 SCIKIT-LEARN CLASSIFIERS

Classifier	**Mean CV-Accuracy**	**SD**
LR	0.441096	0.055715
LDA	0.673217	0.0507
KNN	0.429021	0.054955
CART	0.704918	0.067359
NB	0.096247	0.024918
SVM	0.32655	0.058797
SGD	0.306807	0.068983
RFC	0.7738	0.050682
ETC	0.764709	0.049594
ABC	0.554149	0.049105
BGC	0.784639	0.032237
GBC	0.807576	0.039451
MLP	0.360303	0.085395

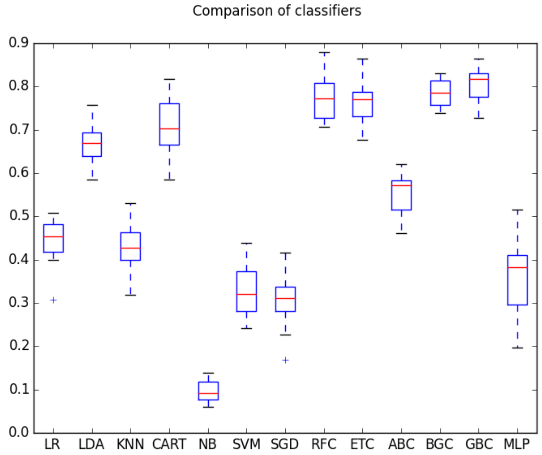

Figure 2. Scikit-Learn Classifiers' Performance on the Current Dataset

C. *Classifier selection (I)*

Regardless of some empirical studies that compare the performance of classifiers' performance with different datasets[18], a common approach is to test multiple models with the current dataset. Thus, the categorised data set has been used to test a total of 13 classifiers provided by Scikit-Learn. At this step, the top four classifiers will be selected using their cross-validation accuracy as the metric. Figure 2 and Table III presents the performance of all tested classifiers measured by their mean cross-validation accuracy, Random Forests (RFC), Extremely Randomised Forests (ETC), Bagging (BGC) and Gradient Boosting (GBC) classifiers show highest CV-accuracy: 77.38%, 76.47%, 78.46%, and 80.76% respectively.

D. *Classifier selection (II)*

Regardless of cross-validation accuracy being a good gross metric for selecting classifiers, it does not provide information about true/false positives/negatives. The tolerance for certain types of error varies from problem to problem. In this case, where cybercrime-related categories of entities are the in the spotlight, minimising false positives while maximising true positives in the tor market, scam, ransomware, mixing, and stolen bitcoin categories, is key in the selection criteria. Tables V, VI, VII, and VIII show the performance of each model and every category. Considering the precision (number of true positives divided by true positives plus false positives) of each model on the specified classes, Table IV shows that overall Gradient Boosting performs the best, followed by Bagging. In all cases, the performance on predicting mixing was poor, while the precision of the other categories range from 50% to 100%.

Table IV
Precision on Cybercrime-Related Categories

Category	***p* RFC**	***p* ETC**	***p* BGC**	***p* GBC**
mixing	0.0000	0.0000	0.0000	0.0000
ransomware	1.0000	0.5000	1.0000	1.0000
scam	0.6000	0.6667	0.6667	0.6667
stolen-bitcoins	0.0000	0.0000	0.0000	1.0000
tor-market	0.8000	1.0000	0.8000	0.8027

Table V
RFC Classification Report

RFC	**precision**	**recall**	**f1-score**	**support**
exchange	0.7042	0.8929	0.7874	56
gambling	0.9048	0.8261	0.8636	23
hosted-wallet	0.0000	0.0000	0.0000	4
merchant-services	0.0000	0.0000	0.0000	4
mining-pool	1.0000	0.7692	0.8696	13
mixing	0.0000	0.0000	0.0000	1
other	0.5714	0.5714	0.5714	7
personal-wallet	0.8817	0.9647	0.9213	85
ransomware	1.0000	0.2000	0.3333	5
scam	0.6000	0.6000	0.6000	5
stolen-bitcoins	0.0000	0.0000	0.0000	2
tor-market	0.8000	0.4444	0.5714	9
avg/total	**0.7822**	**0.8084**	**0.7822**	**214**

Table VI
ETC Classification Report

ETC	**precision**	**recall**	**f1-score**	**support**
exchange	0.6818	0.8036	0.7377	56
gambling	0.7273	0.6957	0.7111	23
hosted-wallet	0.0000	0.0000	0.0000	4
merchant-services	0.0000	0.0000	0.0000	4
mining-pool	0.9091	0.7692	0.8333	13
mixing	0.0000	0.0000	0.0000	1
other	0.3750	0.4286	0.4000	7
personal-wallet	0.8144	0.9294	0.8681	85
ransomware	0.5000	0.2000	0.2857	5
scam	0.6667	0.4000	0.5000	5
stolen-bitcoins	0.0000	0.0000	0.0000	2
tor-market	1.0000	0.4444	0.6154	9
avg/total	**0.7169**	**0.7477**	**0.7222**	**214**

Table VII
BGC Classification Report

BGC	**precision**	**recall**	**f1-score**	**support**
exchange	0.7581	0.8393	0.7966	56
gambling	0.7500	0.7826	0.7660	23
hosted-wallet	0.0000	0.0000	0.0000	4
merchant-services	0.0000	0.0000	0.0000	4
mining-pool	1.0000	0.9231	0.9600	13
mixing	0.0000	0.0000	0.0000	1
other	0.5000	0.5714	0.5333	7
personal-wallet	0.8723	0.9647	0.9162	85
ransomware	1.0000	0.4000	0.5714	5
scam	0.6667	0.4000	0.5000	5
stolen-bitcoins	0.0000	0.0000	0.0000	2
tor-market	0.8000	0.4444	0.5714	9
avg/total	**0.7752**	**0.7991**	**0.7795**	**214**

Table VIII
GBC Classification Report

GBC	**precision**	**recall**	**f1-score**	**support**
exchange	0.7969	0.9107	0.8500	56
gambling	0.8636	0.8261	0.8444	23
hosted-wallet	0.0000	0.0000	0.0000	4
merchant-services	0.0000	0.0000	0.0000	4
mining-pool	0.8571	0.9231	0.8889	13
mixing	0.0000	0.0000	0.0000	1
other	0.2500	0.2857	0.2667	7
personal-wallet	0.8830	0.9765	0.9274	85
ransomware	1.0000	0.4000	0.5714	5
scam	0.6667	0.4000	0.5000	5
stolen-bitcoins	1.0000	0.5000	0.6667	2
tor-market	1.0000	0.5556	0.7143	9
avg/total	**0.8027**	**0.8271**	**0.8056**	**214**

Figure 3 provides a summary of the methodology.

E. Dataset limitation

The dataset utilised in the current research has certain limitations: First, there are some classes that are oversampled compared to some that are substantially undersampled (see Figure 1). While personal-wallets and exchanges have over 200 observations, categories such as stolen-bitcoins or mixing have below 10 observations, which might explain the poor performance of the models when predicting mixing (see Table IV). Second, the variety of classes is limited to the categories that the data provider has successfully identified with their own clustering methodology, hence there is no guarantee that these categories are the only ones in the Bitcoin ecosystem.

F. Alternative approaches

The data provider has developed clustering algorithms and heuristics aiming to identify and categorise as many entities on the Bitcoin Blockchain as possible. However, even though this approach is designed to be as accurate as possible and minimise false positives, it is not scalable. Some algorithms can be applicable to any category, most services have different spending patterns, meaning that for every service there is a new heuristic created. This requires a huge amount of resources in addition to the fact that some services change behaviours over time, and algorithms have to be rewritten.

Within the machine learning domain, an interesting alternative would be applying unsupervised learning clustering. One could argue that the process of building a prototype that leverages unsupervised learning techniques is similar to the one that leverages supervised ones, and that hence the costs are similar. For example, should K-means be the clustering algorithm of choice, it would be interesting to use the same number of categories that our categorised dataset

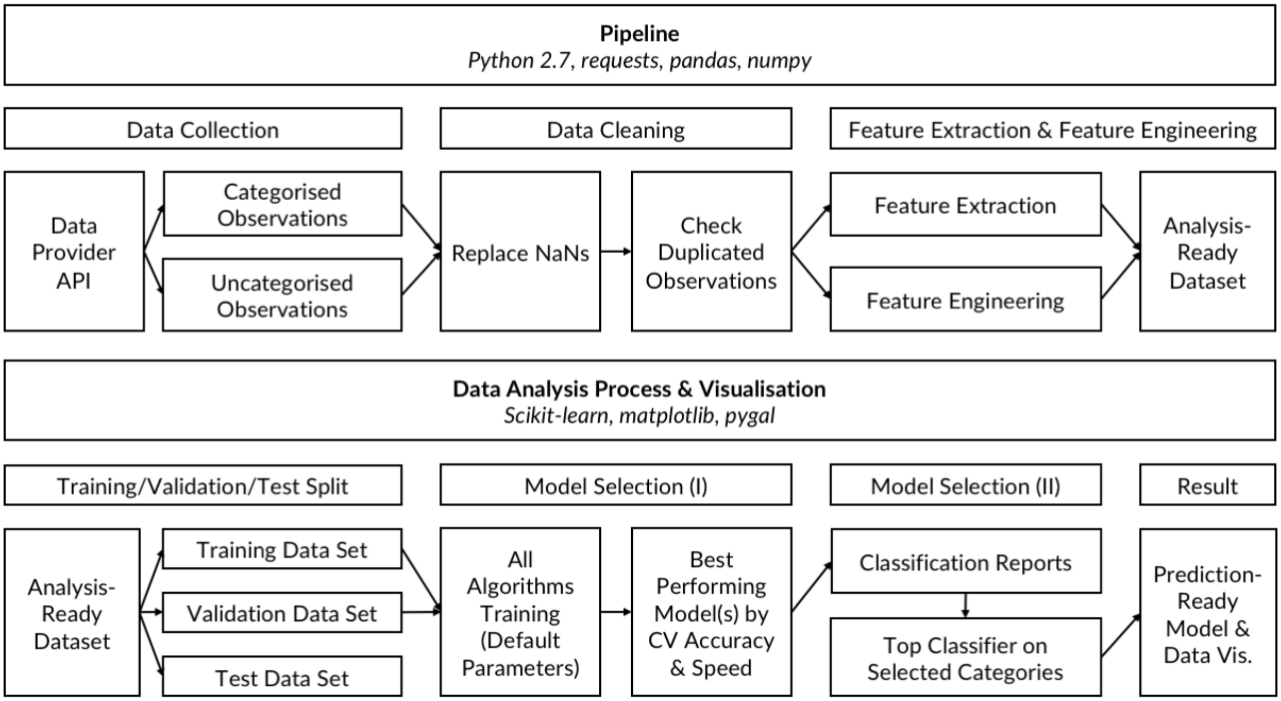
Pipeline
Python 2.7, requests, pandas, numpy
Data Collection
Data Cleaning
Feature Extraction & Feature Engineering
Data Provider API
Categorised Observations
Uncategorised Observations
Replace NaNs
Check Duplicated Observations
Feature Extraction
Feature Engineering
Analysis-Ready Dataset
Data Analysis Process & Visualisation
Scikit-learn, matplotlib, pygal
Training/Validation/Test Split
Model Selection (I)
Model Selection (II)
Result
Analysis-Ready Dataset
Training Data Set
Validation Data Set
Test Data Set
All Algorithms Training (Default Parameters)
Best Performing Model(s) by CV Accuracy & Speed
Classification Reports
Top Classifier on Selected Categories
Prediction-Ready Model & Data Vis.

Figure 3. Pipeline & Data Analysis Process Diagram

provides. Nonetheless, the reasons why it was not chosen for the current research is that it does not fully leverage the available data of the current research, as it would need no categorised observations.

Regardless of the possible approaches, in a scenario where resources such as workforce, computational power, and time are limited, supervised learning remains a viable approach: there are numerous mature and open-source classifiers and it is more scalable than custom heuristics.

V. Results

Considering the process shown in Section IV, the best performing models for classifying the five cybercrime-related categories are Bagging and Gradient Boosting. In this section, we present the outcomes from applying both models to the uncategorised dataset. The results are plotted using three different metrics: number of entities, number of unique addresses, and current balance per category. Table IX provides some statistics representing the coverage of the categorised and uncategorised dataset compared to the entire ecosystem.

Figure 4 is the result of visualising the categorised dataset which includes a total of 854 entities from 12 categories. By looking at the figure, and should the categorised dataset be a representative sample of the Bitcoin ecosystem, one could read that the ecosystem is composed by a relatively small portion of illicit activities-related entities. Figure 5 are 100,000 clusters that, before the classification, had an unknown category. Using the number of entities as metric, the portion of cybercrime-related clusters is 29.81% according to Bagging and 10.95% according to Gradient Boosting. Figure 6 shows the results using number of unique addresses and current balance per category as metrics. 5.79% of addresses, holding a total of 10.02% of coins are the portion of the ecosystem that Bagging has classified as illicit activity-

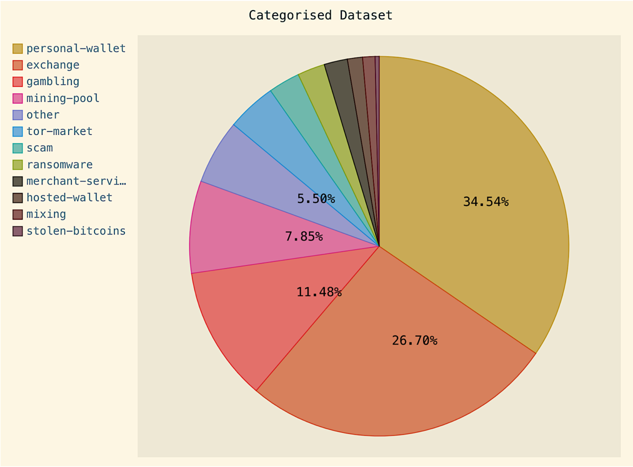

Figure 4. Pie chart of the categorised dataset

Table IX
COVERAGE OF USED DATASETS COMPARED TO THE ECOSYSTEM

Dataset	Clusters	Unique Addresses	%	Total Coins (Balance)	%
Total	?	416,589		16,598,363	
Cat.	854	93,636,522	30.08	519,639.83	3.13
Uncat.	100,000	98,438,675	31.62	4,811,673.72	28.99

related clusters, while Gradient Boosting shows 3.16% and 1.45% respectively.

Tables X and XI show the percentage of cybercrime-related entities in the sample of 100,000 uncategorised clusters.

Table X
RESULTS WITH BAGGING

Bagging					
Number of Clusters		**Unique Addresses**		**Balances**	
Category	**%**	**Category**	**%**	**Category**	**%**
ransomware	19.85	tor-market	5.10	ransomware	9.05
tor-market	9.08	ransomware	0.47	tor-market	0.93
mixing	0.63	scam	0.16	mixing	0.03
scam	0.25	mixing	0.07	scam	0.01
stolen-bitcoins	0.00	stolen-bitcoins	0.00	stolen-bitcoins	0.00
	29.81		5.79		10.02

Table XI
RESULTS WITH GRADIENT BOOSTING

Gradient Boosting					
Number of Clusters		**Unique Addresses**		**Balances**	
Category	**%**	**Category**	**%**	**Category**	**%**
ransomware	5.28	tor-market	2.51	tor-market	0.64
tor-market	4.12	ransomware	0.35	ransomware	0.56
scam	0.92	scam	0.24	scam	0.20
mixing	0.59	mixing	0.05	mixing	0.05
stolen-bitcoins	0.04	stolen-bitcoins	0.00	stolen-bitcoins	0.01
	10.95		**3.16**		**1.45**

The research questions of *how large is the slice of cybercrime-related entities in the Bitcoin ecosystem? How many entities and addresses are related to fraudulent activities? How many Bitcoin are currently held by these entities?* can be answered by looking at tables and figures above:

The 100,000 uncategorised entities comprise of 31.62% of the total unique addresses in the ecosystem and are currently holding 28.99% of the total coins in circulation. The classification shows that the slice of cybercrime-related entities is 29.81% according to Bagging, and 10.95% according to Gradient Boosting using number of entities as metric. When looking at 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 Boosting).

VI. DISCUSSION & FUTURE WORK

Due to the limitations of both the categorised (undersampled classes, volume of observations, variety) and uncategorised datasets (sample of only 100,000, covering 31.62% of unique addresses, and 28.99% of the total coins in circulation), the work can only serve as a proof of concept by providing a very first estimation of how the Bitcoin ecosystem might look like.

Nonetheless, regardless of the dataset limitations, the prototyped model can benefit law enforcement bodies in supporting investigations by narrowing down the list of targets as well as helping prioritise one category over another whenever there is a need. Furthermore, in addition to current investigations, the model can contribute to flagging entities that are likely to be involved in illicit activities. Finally, with further development, it can be used to produce estimations of the ecosystem overtime and reveal past, current and future cybercrime trends.

The scope of the current research project was to provide a rapid prototype. In future work, for the next version of the model, we plan to test additional steps in the methodology to improve the performance of the models: increase the uncategorised dataset (covering 50%, 75%, etc of the total ecosystem); solve the categorised dataset limitations either organically (find more observations) or synthethically increasing the undersampled categories; introduce automatic techniques for dimensionality reduction instead of manual feature extraction and engineering; normalisation of the attributes; and parameter tuning of the algorithms (the current models are using the default parameters as stated in Scikit-Learn's documentation).

VII. CONCLUSION

Bitcoin, a peer-to-peer electronic payment system and digital currency that relies on blockchain technology, has caught the attention of researchers as well as the mainstream media. With the rise of cybercrime activities appearing on headlines, altogether with Bitcoins being involved in scamming, ransomware attacks, tor-markets, and thievery, the cryptocurrency has been commonly associated with only nefarious activities. However, at the time of writing, there has not yet been prior research on the estimation of how the Bitcoin ecosystem looks like nor what types of entities or services can be found. Hence, the purpose of the current research was to answer the following questions: how large is the share of cybercrime-related entities in the Bitcoin ecosystem? How many entities and addresses are related to fraudulent activities? How many Bitcoin are currently held by these entities?. In order to do so, supervised learning techniques were applied.

The methodology had three main components: the data pipeline, built to retrieve the clustered addresses, categorised (a total of 854 observations across 12 different classes) and uncategorised (a total of 100,000 observations), from the

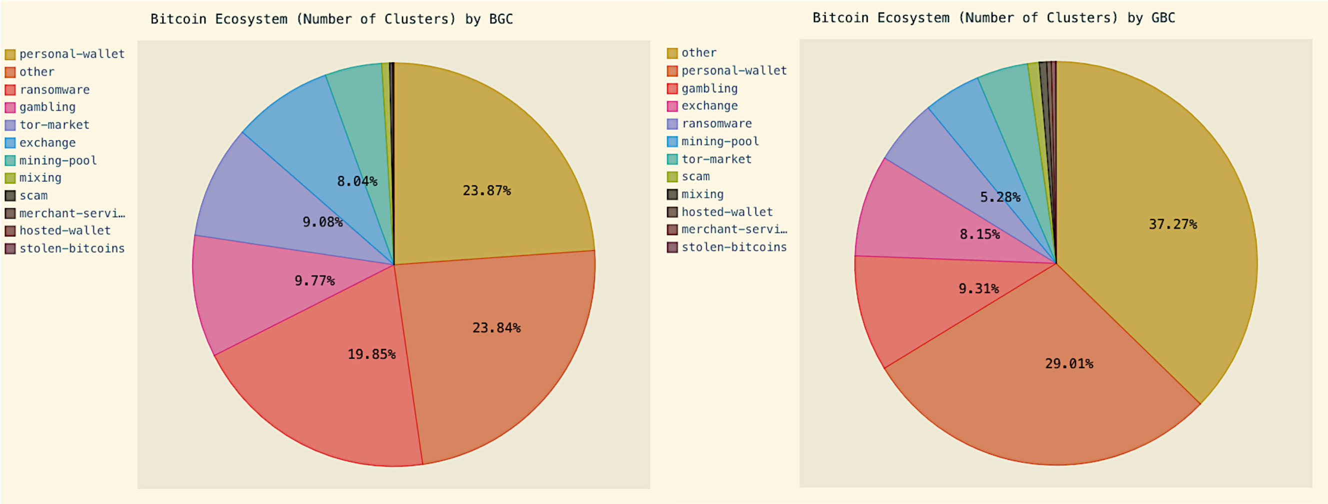

Figure 5. Bitcoin Ecosystem Measured by Number of Entities

data provider, apply preprocessing techniques and produce ready-to-use datasets for the machine learning models; the classifier selection and assessment, where a dozen of different classifiers from Scikit-Learn were tested, resulting in a list of top four classifiers (Random Forest, Extremely Randomised Forests, Bagging, and Gradient Boosting classifiers) by cross-validation accuracy (77.38%, 76.47%, 78.46%, 80.76% respectively), finally top two classifiers by their weighted average precision and precision per class (Bagging and Gradient Boosting Classifier); and the final output production, where classifiers, trained with categorised observations, predict the category of uncategorised observations and charts would be produced with the resulting labels, giving a picture of how the ecosystem looks like in a pseudo-random sample of 100,000 entities that cover 31.62% of unique addresses, and 28.99% of the total coins in circulation.

Due to the dataset limitations, the outcomes serve as a limited estimation of the Bitcoin ecosystem. Nonetheless, the prototype can still benefit law enforcement bodies during analysis and investigations filtering the list of suspicious addresses not only by number, but also by category depending on the priorities of investigators. Moreover, besides current investigations, the model can flag entities that are more likely to be related to cybercriminal activities, as a form of prevention. Finally, with further development of the model and a refined methodology, it can be used to produce estimations of the entire ecosystem and reveal past, current and future cybercrime trends.

References

[1] S. Nakamoto, *Bitcoin: A peer-to-peer electronic cash system*, 2008.

[2] A. Cuthbertson, *Bitcoin now accepted by 100,000 merchants worldwide*, International Business Times. IBTimes Co., Ltd. [Accessed November 20, 2015]

[3] G. Hileman and M. Rauchs, *Global cryptocurrency benchmarking study*, Cambridge Centre for Alternative Finance, 2017.

[4] M. A. Harlev, H. Sun Yin, K. C. Langenheldt, R. Mukkamala, and R. Vatrapu *Breaking Bad: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning*, Proceedings of 51st Hawaii International Conference on System Sciences (HICSS), 2018.

[5] FBI: Cyber Intelligence Section and Criminal Intelligence Section, *Bitcoins Virtual Currency: Unique Features Present Challenges for Deterring Illicit Activity*, FBI. 24 April 2012 [Accessed November 2, 2014].

[6] Symantec Corporation, *Ransomware and Businesses 2016. Symantec Corporation*, 2016.

[7] Symantec Corporation, *Internet Security Threat Report: Volume 22*, 2017.

[8] K. Cabaj, P. Gawkowski, K. Grochowski, and D. Osojca *Network activity analysis of CryptoWall ransomware. Przeglad Elektrotechniczny*, 3rd ed. Przeglad Elektrotechniczny, 91(11), 201-204, 2015.

[9] K. Liao, Z. Zhao, A. Doup, and G. J. Ahn, *Behind closed doors: measurement and analysis of CryptoLocker ransoms in Bitcoin*, 3rd ed. Electronic Crime Research (eCrime), 2016 APWG Symposium on (pp. 1-13), IEEE, 2016.

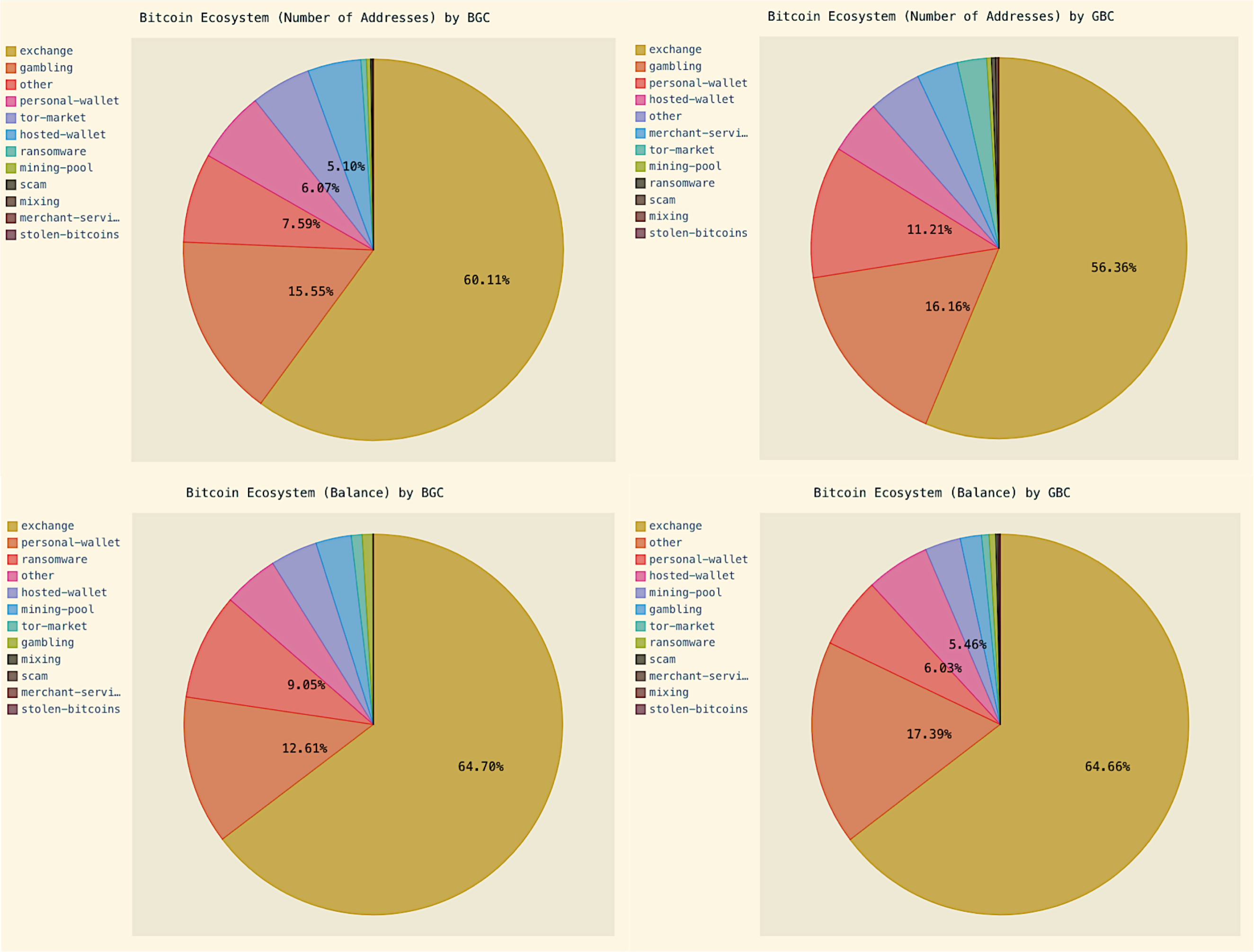

Figure 6. Bitcoin Ecosystem Measured by Addresses and Balance

[10] N. Christin, *Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace*, 3rd ed. Proceedings of the 22nd international conference on World Wide Web (pp. 213-224), ACM, 2013.

[11] M. Mser, *Anonymity of bitcoin transactions*, 3rd ed. Mnster bitcoin conference (pp. 17-18), 2013.

[12] F. Reid and M. Harrigan, *An analysis of anonymity in the bitcoin system*, 3rd ed. Security and privacy in social networks (pp. 197-223). Springer New York, 2013.

[13] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage*A fistful of bitcoins: characterizing payments among men with no names*, 3rd ed. Proceedings of the 2013 conference on Internet measurement conference (pp. 127-140), ACM, 2013.

[14] M. Spagnuolo, F. Maggi, and S. Zanero*Bitiodine: Extracting intelligence from the bitcoin network*, 3rd ed. International Conference on Financial Cryptography and Data Security (pp. 457-468). Springer, Berlin, Heidelberg, 2014.

[15] D. Reid and A. Shamir, *Quantitative analysis of the full bitcoin transaction graph*, 3rd ed. International Conference on Financial Cryptography and Data Security (pp. 6-24). Springer, Berlin, Heidelberg, 2013.

[16] J. Hirshman, Y. Huang, and S. Macke*Unsupervised approaches to detecting anomalous behavior in the bitcoin transaction network*, 3rd ed. Technical report, Stanford University, 2013.

[17] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*, 3rd ed. Princeton University Press, 2016.

[18] R. Caruana and A. Niculescu-Mizil, *An empirical comparison of supervised learning algorithms*, 3rd ed. International Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM, 2006.