Breaking Bad: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning

*Mikkel Alexander Harlev[1], Haohua Sun Yin[1], Klaus Christian Langenheldt[1],
Raghava Rao Mukkamala[1,2], and Ravi Vatrapu[1,2]
[1]Centre for Business Data Analytics, Copenhagen Business School, Denmark
[2]Westerdals Oslo School of Arts, Comm & Tech, Norway
{miharlev, awasunyin, k.langenheldt}@gmail.com, {rrm.itm, rv.itm}@cbs.dk

Abstract

Bitcoin is a cryptocurrency whose transactions are recorded on a distributed, openly accessible ledger. On the Bitcoin Blockchain, an entity's real-world identity is hidden behind a pseudonym, a so-called address. Therefore, Bitcoin is widely assumed to provide a high degree of anonymity, which is a driver for its frequent use for illicit activities. This paper presents a novel approach for reducing the anonymity of the Bitcoin Blockchain by using Supervised Machine Learning to predict the type of yet-unidentified entities. We utilised a sample of 434 entities (with ≈ 200 million transactions), whose identity and type had been revealed, as training set data and built classifiers differentiating among 10 categories. Our main finding is that we can indeed predict the type of a yet-unidentified entity. Using the Gradient Boosting algorithm, we achieve an accuracy of 77% and F1-score of ≈ 0.75. We discuss our novel approach of Supervised Machine Learning for uncovering Bitcoin Blockchain anonymity and its potential applications to forensics and financial compliance and its societal implications, outline study limitations and propose future research directions.

1. Introduction

Bitcoin is a cryptocurrency and a global distributed payment system on which transactions are facilitated through a peer-to-peer network. Bitcoin was first described in 2008 [1] and ever since has attracted the attention of the research community from diverse academic fields [2] [3] [4] and gained mainstream popularity due to its unique characteristics, such as the absence of centralised control and an assumed high degree of anonymity.

Because of Bitcoin's comparably high level of anonymity, it has been labelled as the go-to currency for illicit activity. The shutdown of the drug market Silk Road[1] provides the most well-known example in this context (see [5] for an analysis of Silk Road). Moreover, there have been articles and reports [6–8] stating that Bitcoin has been used for terror financing, thefts, scams and ransomware. This is why financial regulators, law enforcement, intelligence services and companies who transact on the Bitcoin Blockchain have become wary observers of technical developments in, economic issues with, and the societal adoption of the cryptocurrency Bitcoin [2–4]. For companies, interacting with high-risk counterparties on the Bitcoin Blockchain may yield negative consequences, either because of legal obligations (such as anti-money laundering procedures) or reputational risks. For governments, the fact that Bitcoin is used to carry out money-laundering, terror financing or cybercrime poses a considerable problem. In such cases, uncovering the anonymity of the parties would be legally permissible and ethically desirable - but technically infeasible, according to popular belief about the robustness of the Bitcoin Blockchain's anonymity.

However, previous research [9] [10] has demonstrated that it is indeed possible to cluster together Bitcoin addresses and link such *clusters* to real-world identities. These research findings go against the widely-held belief that users' identities are protected when using Bitcoin. Our work builds upon and extends this area of research, investigating the true level of Bitcoin's anonymity. Knowing that Bitcoin addresses can be clustered, identified and categorised, we investigate if it is possible to reveal (to some extent) the identity of users or organisations on the Bitcoin Blockchain using a Supervised Machine Learning approach.

Problem Formulation & Research Question

For this research paper, we collaborated with the Bitcoin analysis company *Chainalysis* [11], which

*The first three authors contributed equally for the first authorship.

[1] https://en.wikipedia.org/wiki/Silk_Road_(marketplace)

URI: http://hdl.handle.net/10125/50331
ISBN: 978-0-9981331-1-9

will be referred to as the *data provider* in the rest of the paper. The data provider has clustered, identified and categorised a substantial number of Bitcoin addresses manually or through a variety of clustering techniques (sec. 4). However, the vast majority of clusters on the Bitcoin Blockchain remain uncategorised. Our research aims to find out if we can predict that a yet-unidentified cluster belongs to one of the following pre-defined categories: *exchange*, *gambling*, *hosted wallet*, *merchant services*, *mining pool*, *mixing*, *ransomware*, *scam*, *tor market* or *other*. We recognise the fact that there are additional cluster types participating in the Bitcoin economy, but the scope of our research will be limited to said categories, as those are the categories provided by the data provider. At the time of writing, to the best of our knowledge, there has not yet been any research utilising such enriched data. Furthermore, alternative data sources, offering more identified clusters than the data provider's, remain unknown. It must be noted that this work prioritises demonstrating the feasibility of using Supervised Learning to reduce Bitcoin Blockchain's anonymity over ensuring complete data reliability, as the clustering methodology of the data provider is proprietary and confidential. Based on the above discussion, our research question is as follows:

> *To what extent can we predict the category of a yet-unidentified cluster on the Bitcoin Blockchain?*

The outline of this paper is as follows; in sec. 2, a brief overview of related work is provided and Sec. 3 presents the theoretical foundations. The methodology is presented in Sec. 4 and Sec. 5 provides an overview of the results and their substantive interpretation. Finally Sec. 6 discusses implications of our findings, and considers future work.

2. Related Work

The research on Bitcoin anonymity ranges from explaining the technological workings of Bitcoin and Blockchains to analysing the limitations and challenging the features of the technology, especially its pseudo-anonymity. Many researchers explored the limits of the anonymity of Bitcoin by applying heuristic approaches or statistical methods. For instance, clustering the Bitcoin addresses by mapping the network, analysing the traffic and complementing it with external pieces of information was explored in [10]. Another reserch work [12] analysed the Bitcoin system by simulation experiments by replicating the behaviours and transactions of the Bitcoin Blockchain. By using categories created from interactions with certain services, multiple addresses belonging to the same user were clustered together in [9], where as [13] has shown that it is possible to identify behavioural patterns of different types of users by creating Bitcoin Blockchain transaction graphs and by analysing its statistical properties. By gathering real-time transactions over a time frame, [14] developed heuristics aiming to cluster and reveal the ownership behind the Bitcoin addresses and IP addresses. Finally, another work [15] developed a graph analysis framework that uses both Bitcoin Blockchain and data scraped from online forums and social media platforms that belong to crypto currencies.

In the domain of Data Mining and Machine Learning, primarily the research focus was mainly to contribute to more effective crime investigation. Some of the notable research include detecting suspicious transactions by applying Unsupervised Learning techniques [16] and building guidelines for data collection and a framework for data processing and extraction, specifically against fraud, false transactions and money theft [17]. Another line of research is focused on public forums data, constructing a network topology, and analysing a criminal cluster such as CryptoLocker, a known family of ransomware in the Bitcoin economy [18]. The work in [19] focussed on applying clustering algorithms such as multi-input heuristics for the data analysis.

Many researchers also focussed on the flaws of the Bitcoin Blockchain and explored alternative cryptocurrencies as well as proposals for improvements / new methods to bring anonymity to its users. Some of the research works explored in-depth investigation on Bitcoin's technological workings, showing its technological flaws and consequent suggestions on how to address them [20], a protocol that enables anonymous payments in Bitcoin and other currencies, which relies on technology commonly used by mixing services [21]. In this regard, an important research contribution is on building an alternative to Bitcoin named Zerocash with zero-knowledge proofs [22] and also privacy-enhancing overlays in Bitcoin from a theoretical perspective [23].

For the majority of the aforementioned research, the researchers collected the Blockchain data on their own, crafted their own categories and extracted their own intelligence. Because the data provider supplied us with rich data of already collected, clustered, categorised and identified addresses, we were able to focus on the data analysis process from the very start. In contrast to the existing research, our analysis approach focused on utilising Supervised Machine Learning methods to categorise the yet-unidentified entities. To the best

of our knowledge, there is no other research work that focused on de-anonymising the Bitcoin Blockchain using Supervised Machine Learning techniques.

3. Conceptual Framework

In this section, we first present the basic concepts of the technological workings of the Bitcoin Blockchain and following present the clustering methodology applied to cluster Bitcoin addresses. Finally, we will describe the categories used to label the clusters.

3.1. Basic Concepts

In order to transact on the Bitcoin Blockchain, a user receives a pseudonym, an *address*. A user may create as many such addresses as desired to enhance anonymity (it is advised as best practice to create a new address for each new transaction [24]). A *transaction* primarily consists of four main elements: 1) Transaction hash value, 2) Address of the sender, 3) Address of the receiver and 4) Amount. The Bitcoin Blockchain holds additional data, which will be discussed later in this paper. Furthermore, a transaction may involve more than one input and/or output address, making it challenging to link multiple transactions to one person. This manifests through, for example, the so-called *change address*: Each transaction initially draws all Bitcoin from a users account balance, then sends one part of the amount to the desired receiver address and the remaining part (the change) to a change address. The change address can be same as the original sender address, but it is a best practice to create a new change address for each transaction. Subsequently, to approve a transaction, the sender must use the corresponding *private key* to sign a transaction. The transaction is then sent to the network, collected into *blocks* along with other transactions, after being verified, then accepted into the Blockchain by the consensus of all peers. Finally, the transaction is broadcasted to the network and becomes publicly visible.

The power of the Bitcoin Blockchain lies in the fact that each and every interaction is recorded on an immutable, publicly accessible ledger. This makes Bitcoin well-suited for high-trust applications (e.g. money transfer) that traditionally require a reliable intermediary (e.g. clearing houses) to validate transactions. To preserve the anonymity of Bitcoin users, their identities are hidden behind an *address*, also referred to as *public key* or *pseudonym*. This pseudonym cannot directly be linked to the real-world identity of the user. The problem of Bitcoin's architecture is that once a pseudonym gets linked to a real-world identity, it effectively reveals all transactions undertaken by that pseudonym, with no way of deleting the corresponding transaction history. Such identity-revealing linking can occur either through voluntary disclosure (e.g. when a vendor publicises its address in order to receive Bitcoin from its customers), or through involuntary disclosure (like data leakages, addresses taken from court documents or data exchange partnerships between Bitcoin companies). Such clear-cut identification is however seldom possible. However, there is a variety of methods to effectively narrow down the scope of who could own a given Bitcoin address. Like Reid et al. (2013) [10] have found, it is possible to link the *change address* of a transaction back to the initial user. Further, it is possible to cluster together individual addresses that are controlled by the same person using different clustering techniques [25]. Moreover, it is even possible to map IP addresses to Bitcoin addresses. Our approach is to narrow down the scope of possible owners of a cluster by predicting the category of a yet-unidentified cluster using supervised Machine Learning approaches.

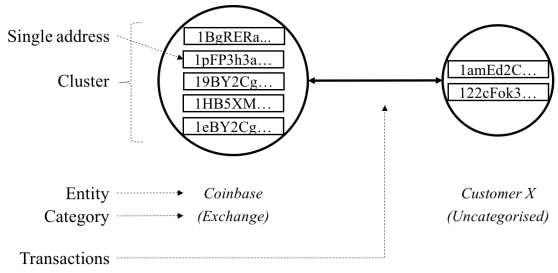

Figure 1. Anatomy of a Bitcoin Cluster

As shown in Fig. 1, an *entity* is defined as a person or organisation believed to be in control of a single or multiple addresses. A *cluster* is defined as a group of addresses controlled by one entity. Corresponding to the entity's main activity or nature, it can be given a *category*). The data provider currently assumes that every entity can only belong to one of a category at a time, which means that the categories are mutually exclusive. Figure 1 shows an example of two Bitcoin clusters (*Coinbase* and *Customer X*), where one can observe individual Bitcoin addresses are grouped into a cluster, pertaining to an entity. The first entity (*Coinbase*) is labeled with a category label of *Exchange*. An Exchange allows their customers to trade Bitcoins for fiat currencies, whereas the other entity is labeled *Uncategorised*, meaning the cluster has not yet been identified (i. e. it has not yet been linked to a real-world identity).

3.2. Clustering Methodology

The transactional data used by the data provider is publicly available to everyone and can be retrieved from the Bitcoin Blockchain without any cost. However, the data used in this research has been enriched through various data processing techniques, providing us with addresses that have already been clustered, identified and categorised. As defined earlier, a *cluster* is a collection of Bitcoin addresses that are estimated to be controlled by a single entity. Clusters are identified by the data provider through different means, as follows,

- *Co-spend clustering*: A co-spend cluster is estimated due to several addresses all contributing inputs to a single transaction. Suppose, that a user sends a Bitcoin to a merchant, with 0.4 Bitcoin coming from one address and 0.6 Bitcoin from another. Prior to this transaction the two sending addresses would appear to be two separate entities. However, after the transaction takes place we can conclude that there is only one entity behind the transaction as both private keys would need to be present to sign the transaction as valid. Not only are the addresses thus linked to this transaction, but all previous and future transactions involving those addresses are now linked, too.
- *Intelligence-based clustering*: In this type of clustering, information is gathered from outside the Blockchain. The data sources from which information is gathered, include but are not limited to: data leaks, court documents, data partnerships, exchanges that share their addresses and manual merges due to services changing wallets.
- *Behavioural clustering*: As part of this clustering, patterns in the timing or structure of transactions will be utilised to identify a specific wallet. Basically a *wallet* is nothing but a Bitcoin equivalent of a bank account, where users store and transact their Bitcoins. There can be a software wallet (like an application installed by the users on their devices) or a web / hosted wallet, which is normally hosted and maintained securely by a third party provider. Behavioural clustering can be used to cluster and relate the Bitcoin addresses to known hosted services or even to a specific wallet software.

The data provider sends at least one transaction to every cluster before categorising it and tracks the moving funds to ensure that the clustering is error-free. Finally, considering that the data is used in law enforcement and financial compliance, the clustering algorithms and heuristics are designed and reinforced to minimise false positives, as errors could cause serious repercussions.

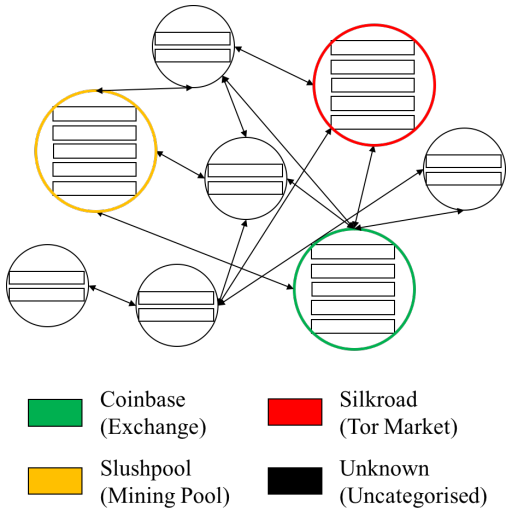

Figure 2. Visualisation Network of Different Categories

3.3. Cluster Categories

The Bitcoin addresses that were clustered together using above mentioned clustering techniques are further labelled with different category labels assigned by the data provider. These category labels can range from non-suspicious activities such as *exchanges* to high-risk categories, such as *ransomware* and so on. The below are the categories used in this work for predicting an yet-unidentified entity.

- *Exchange*: Entities that allow their customers to trade fiat currencies for Bitcoins
- *Hosted-Wallet*: Trusted entities that offer Bitcoin storage as a service
- *Merchant Services*: Entities that offer solutions to businesses in order to facilitate the adoption of Bitcoins as a payment method for their customers
- *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
- *Mixing*: Entities that apply techniques to reduce the traceability of their clients' transactions as a service
- *Gambling*: Entities that offer gambling services
- *Scam*: Entities that deceive their customers by pretending to provide a service in order to steal their Bitcoins
- *Tor Market*: Marketplaces primarily facilitating trading of illegal goods like narcotics, stolen credit cards, passports etc. These sites are only accessible on the deep web through e.g. the TOR-browser
- *Ransomware*: Entities that are utilising the Bitcoin Blockchain as a medium of exchange to receive

ransom fees
- *Other*: Entities that have been identified but do not belong to any of the nine categories mentioned above, for example WikiLeaks' donation address

Given the above described clustering techniques, the consequently revealed cluster identities and their corresponding categories, we can illustrate a network of identified clusters.

4. Methodology

In this section, we will first describe the dataset and its primary characteristics, then we will discuss the choice between seven Supervised Machine Learning algorithms and build a classifier in order to predict the category of yet-unidentified clusters. We will also discuss the need for using over-sampling to deal with class imbalance problems of under-represented categories and finally conclude with the dataset's limitations.

4.1. Data Analysis Process

As mentioned before, the dataset used in this research was provided by the company *Chainalysis* [11], which is specialised in Bitcoin Blockchain analysis. The dataset primarily contains transactional data, containing details about every single transaction an entity has participated in, such as the timestamp, the value sent or received in Bitcoins and USD, or the counter-party of the transaction. In addition to this, the dataset also contains the characteristics of each cluster and in some cases the categories have already been identified.

As shown in Table 1, the total dataset used in this research contains approximately 200 million transactions pertaining to 434 unique clusters. The number of transactions per cluster varies significantly, ranging from a low number (≤ 10) to several million transactions. Additionally, table 2 illustrates the number of transactions for each category. More specifically, the *average, median, minimum* and *maximum* number of transactions are shown, based on the observations within the respective category. For each transaction, there are several describing attributes, as shown in Table 3. This information is utilised to capture the cluster's behaviour using time-series analysis. To describe the behaviour of a cluster in a way that can be fed to a Supervised Machine Learning algorithm, we extracted a set of features from the original input variables (see Table 3) for each identified cluster. Apart from the extracted features, we engineered additional features such as the number of transactions, their mean and standard deviation, the cluster lifetime, a cluster's exposure to specific other clusters, and so forth. The resulting feature space consists of a total of 76 features.

Total number of transactions	198,097,356
Number of unique Clusters	434
Average number of transactions per cluster	456,445.52
Lowest number of transactions in a cluster	7
Highest number of transactions in a cluster	26,937,988

Table 1. Dataset Description

Category	Avg. TRX	Median TRX	Min. TRX	Max. TRX
Exchange	453116	34313	9	26937988
Gambling	337255	17823	62	13084669
Hosted Wallet	1077015	112562	11531	8657104
Merchant Services	568629	122627	1412	3144731
Mining Pool	786820	97256	914	11941952
Mixing	6490196	122012	267	25716495
Other	83424	12074	476	659181
Ransomware	2225	1799	68	8038
Scam	21277	18835	759	68054
Tor Market	281034	17874	70	3194191

Table 2. Number of transactions (TRX) per category

Feature Name	Description
TRX Date	Timestamp of the transaction
TRX BTC Received	Amount of BTC received (Blank if the entity is the sender)
TRX BTC Sent	Amount of BTC sent (Blank if the entity is the receiver)
TRX USD Value	The equivalent USD amount at the point in time
TRX Peer Category	The entity type of the counterparty (e.g. exchange or tor-market)
CP BTC Sent	The total BTC amount sent to a given cluster
CP BTC Received	The total BTC amount received from a given cluster
CP TRX Output Count	The total number of transactions conducted with the given cluster
CP BTC Flow	The numerical value of received BTC minus sent BTC with the given cluster

Table 3. List of Original Variables of the Dataset

As shown in Figure 3, the first phase in the data analysis process is the data preparation, which contains data preprocessing and feature extraction as the main processing steps, in order to transform the dataset to be readable by the Machine Learning algorithms. The data analysis phase consists of three main steps. In the first step, we selected and trained a set of multi-class classifiers using their default parameters. This step provided a preliminary evaluation of which algorithms could be suitable for the problem at hand. In the second step, we tuned the hyper-parameters for each model using cross-validated random search. We assessed the performance of each algorithm after training each model with their respective set of optimal parameters. Finally, in order to compensate for the class imbalance, we used SMOTE (Synthetic Minority Over-Sampling Technique [26]) to oversample the two minority classes

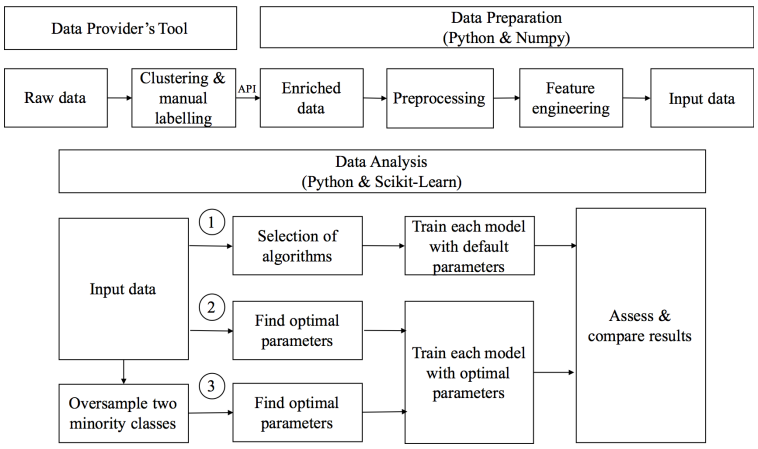

Figure 3. Data Preparation and Analysis Process

hosted-wallet and *mixing*. After parameter tuning with the oversampled data, each model was trained and assessed again using Scikit-learn, [27] a machine learning library. After assessing all the results, a classification report and ROC curve of best performing algorithm were produced.

4.2. Supervised Machine Learning Algorithms

For the analysis of Bitcoin transaction data, we used the following Supervised Machine Learning algorithms, which are popular choices for this type of problem [28].

- k-Nearest Neighbours,
- Random Forests,
- Extra Trees,
- AdaBoost,
- Decision Trees,
- Bagging Classifier,
- Gradient Boosting.

Regarding the choice of algorithms, we have excluded linear models and Support Vector Machine, as our dataset includes a variety of collinear variables which may increase the variance of the coefficient estimates, sensitivising the model to minor changes [29]. Parameter tuning was undertaken using 3-fold cross validation due to the scarcity of known clusters (n = 434). Therefore, using a traditional train-test-validation split would bear the risk of making the performance too dependent on a specific subset of training data, waste data and inhibit predictive ability [30]. As for hyper-parameter optimisation, random search was utilised with 1000 iterations, since it is empirically and theoretically more effective than grid search, as it allows the testing of a broader value spectrum for each parameter, as well as being less likely to waste effort on irrelevant hyper-parameters, given the same amount of iterations [31]. A detailed description of the chosen algorithms could be included due to space constraints.

Class Imbalance: Our dataset contains two minority and under-represented classes *hosted-wallet* and *mixing*. There are two reasons why certain classes are under-sampled. First, some categories (e.g. *mixing*) wish to remain unidentifiable due to the nature of their activities and thus apply privacy-enhancing schemes. For example, they obfuscate transactions through so-called *peeling chains*: a *mixing service* takes a customers' deposits and moves it to one single address. Then, it starts sending very small amounts from this address to different services and the remaining coins (the change) to a new change address; this process is repeated until the very last coin has been spent. This creates dozens or even hundreds change addresses, obfuscating the actual origin of a transaction, making it hard to identify and cluster addresses. Second, the data provider prioritises some categories over others, depending on their customers' needs and cybercrime trends, which is why classes such as *hosted-wallet* have less observations. Clustering and identifying entities is an ongoing process, hence the data provider increases the number of categorised entities as time passes.

To deal with this class imbalance problem, we used the SMOTE [26] method to balance out the under-represented classes. This approach constructs synthetic samples of the minority classes to improve the utility of imbalanced datasets. It has been shown that by increasing a classifier's sensitivity to the minority class through increasing its sample size, the prediction model can achieve a better performance [26, 32]. The method had been applied with a ratio of 0.075 for the two classes with the fewest observations: *mixing* and *hosted-wallet*. This means that synthetic samples had been generated to the point where the minority classes reach 7.5% of the amount of samples in the majority class. Since the

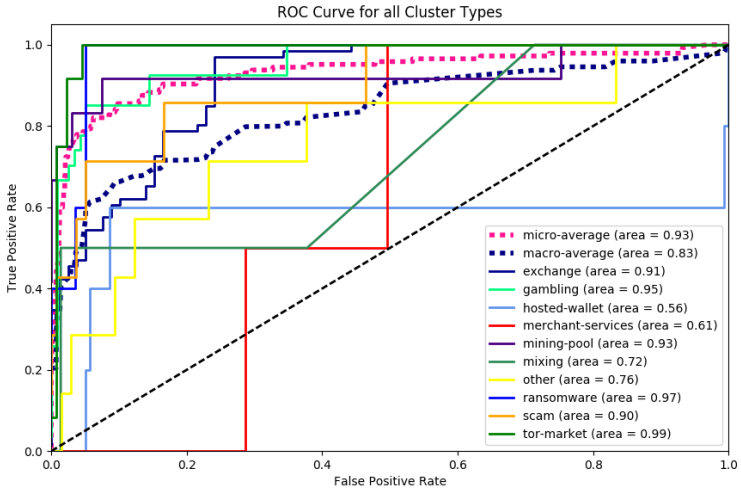

Figure 4. ROC Curve with GBC

majority class *exchange* consists of 203 samples, the two minority classes ended up having 15 samples each. As we are using 3-fold cross validation, assessing the true contribution of SMOTE is limited.

Dataset Limitations: Currently, the data used to train the prediction model does not include all of the data that is inherently available on the Bitcoin Blockchain. This applies to the *transaction fee* that is associated with the transaction priority, the amount of signatures used to sign a transaction, the related IP address or the transaction size and the number of confirmations, among others. Therefore, additional features could be extracted in order to increase the performance of our predictions. Additionally, the amount of clusters used to train the prediction model is limited to those that have already been categorised by the data provider. While we do have more than 400 categorised clusters, a larger sample size could potentially allow to discover more categories, as well as increase the number of examples for each of the already-defined 10 categories, thereby improving the performance of the model. Finally, given a larger sample size, the methodology could be improved by utilising a test sample, not seen by the classifier, to accurately justify the final results of the model.

5. Results

We split up the dataset into one with SMOTE (i.e. adjusted for class imbalance) and one without, and subsequently applied the data analysis process to both datasets (Figure 3). Out of the seven algorithms, Random Forest, Bagging and Gradient Boosting provided the best-performing models for both with and without oversampling. The achieved accuracies are 73% for Random Forest, 74% for Bagging and 77% for Gradient Boosting Classifier (GBC). Performance measures of three best classifiers are provided in Table 4 and others are skipped due to space constraints.

Algorithm	Acc.	Prec.	Recall	F1-score	Support
Random Forest	0.73	0.71	0.71	0.67	434
Bagging	0.74	0.73	0.76	0.72	434
Gradient Boosting	0.77	0.74	0.77	0.75	434
Gradient Boosting with SMOTE	0.78	0.75	0.78	0.76	451

Table 4. Performance of the Three Best Classifiers

5.1. Results without SMOTE

Gradient Boosting proved to be the best-performing algorithm. The corresponding classification report is illustrated along with the set of hyper-parameters utilised (as shown in fig. 6 and Table 5). The low sample size resulted in the depicted jagged lines as shown on the ROC curve with GBC (see Figure 4). Typically, ROC curves are smoother (similar to the micro and macro average curves as shown on the same Figure). It also shows that the model struggles when predicting the categories *hosted wallet* and *mixing*, which could be explained by the considerably lower sample size. Table 5 shows *precision*, *recall*, *F1-score* and *support* obtained from classification with GBC.

As for *merchant services*, regardless of the sample size being larger than *hosted wallet* and *mixing*, one could interpret that the sample size is too low and that the observations are not differentiated enough from the other 9 classes. Additionally, the model has difficulty

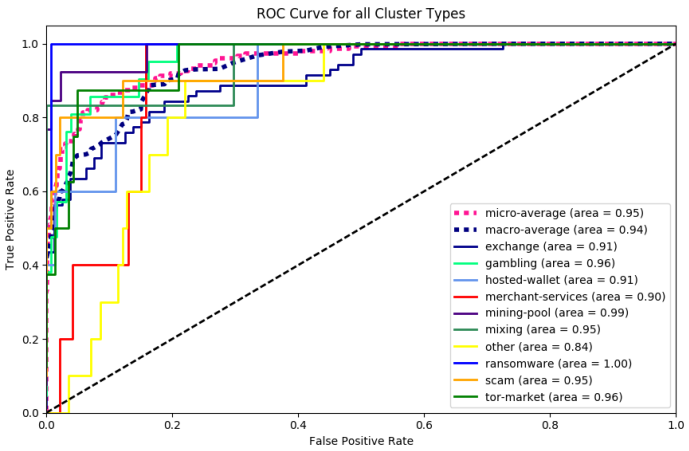

Figure 5. ROC Curve with GBC & SMOTE

```
GradientBoostingClassifier(criterion='friedman_mse',
  ↪  init=None, learning_rate=0.1, loss='deviance',
  ↪  max_depth=13, max_features='auto',
  ↪  max_leaf_nodes=None, min_impurity_split=1e-07,
  ↪  min_samples_leaf=11, min_samples_split=2,
  ↪  min_weight_fraction_leaf=0.0, n_estimators=38,
  ↪  presort='auto', random_state=None, subsample=1.0,
  ↪  verbose=0, warm_start=False)
```

Figure 6. Parameters for Gradient Boosting Classifier

```
GradientBoostingClassifier(criterion='friedman_mse',
  ↪  init=None, learning_rate=0.1, loss='deviance',
  ↪  max_depth=8, max_features='sqrt', max_leaf_nodes=None,
  ↪  min_impurity_split=1e-07, min_samples_leaf=15,
  ↪  min_samples_split=2, min_weight_fraction_leaf=0.0,
  ↪  n_estimators=56, presort='auto', random_state=None,
  ↪  subsample=1.0, verbose=0, warm_start=False)
```

Figure 7. Parameters for GBC with SMOTE

Category	Precision	Recall	F1-score	Support
Exchange	0.79	0.94	0.86	201
Gambling	0.74	0.83	0.78	89
Hosted Wallet	0.25	0.11	0.15	9
Merchant Services	0.00	0.00	0.00	13
Mining Pool	0.96	0.84	0.90	31
Mixing	0.50	0.25	0.33	4
Other	0.43	0.15	0.22	20
Ransomware	0.91	0.77	0.83	13
Scam	0.68	0.59	0.63	22
Tor Market	0.79	0.59	0.68	32
Avg / Total	0.74	0.77	0.75	434

Table 5. Classification Report with GBC

Category	Precision	Recall	F1-score	Support
Exchange	0.76	0.94	0.84	201
Gambling	0.77	0.80	0.78	89
Hosted Wallet	0.88	0.47	0.61	15
Merchant Services	0.00	0.00	0.00	13
Mining Pool	0.96	0.77	0.86	31
Mixing	1.00	0.93	0.97	15
Other	0.17	0.05	0.08	20
Ransomware	0.85	0.85	0.85	13
Scam	0.68	0.59	0.63	22
Tor Market	0.96	0.72	0.82	32
Avg / Total	0.75	0.78	0.76	451

Table 6. Classification Report with GBC & SMOTE

predicting observations from the *other* category, which can be explained by the fact that the *other* category encapsulates a multitude of categories. In order to deal with class imbalance and to achieve a better performance, SMOTE has been applied to oversample *hosted-wallet* and *mixing*. The results can be seen in the Figure 5.

5.2. Results with SMOTE

Applying SMOTE to compensate for the class imbalance improved the performance of predicting under represented classes *mixing* and *hosted wallet*. The overall performance of the model was increased to an accuracy of 78% as shown in Table 4. Additionally, the corresponding classification report showing *precision*, *recall*, *F1-score* and *support* obtained from classification with GBC is given in the Table 6. However, even though the overall results with over-sampling are slightly better than the results with the original dataset, the increase could potentially be a product of overfitting, hence we discard the slightly improved results generated with SMOTE. In summary, from all the tested algorithms, Gradient Boosting provides the best performance: The ROC-curve displays a micro- and macro-average of respectively 0.93 and 0.83. Additionally, an accuracy of 77%, a precision of 74%, a recall of 77%, and an F1-score of 75% were achieved as shown in classification report (Table 5).

6. Discussion

In this research work, we have demonstrated a novel method to categorise yet-unidentified clusters on the Bitcoin Blockchain using Supervised Machine learning. Our results show that we can predict the category of a yet-unidentified cluster on the Bitcoin Blockchain with a 77% accuracy (and an F1-score of 0.75) using Gradient Boosting Classifier. Admittedly, the research is limited by the sample size of 434 observations. As shown previously, our model struggles when predicting classes with low number of observations, such as *mixing* and *merchant services*. Furthermore, accuracy could be improved by enhanced feature engineering, for example by using automated time-series feature extraction. Additionally, one could consider alternative approaches to our analysis, such as transforming the problem into a binary classification problem and only predicting one specific class (e.g. non-scam/scam), reducing randomness and allowing to choose from a broader set of algorithms. While our set of chosen algorithms are computationally expensive, speed is not an issue, as we don't target real-time prediction, thus we are outweighing the computation time of the model by the higher performance and accuracy of the model.

The results show that it is possible to categorise yet-unidentified clusters which means that one could reveal the category of a significant portion of entities on the Bitcoin Blockchain, which further challenges popular beliefs about Bitcoin's true anonymity. With regard to practical applications, our approach could potentially contribute to crime investigation, e.g. by flagging suspicious entities such as ransomware or scams. Finally, due to some countries' regulations, a company that transacts on the Bitcoin Blockchain might be obliged to prove that the received money had not been involved in illicit activities. For such compliance tasks, our current research paves the way towards identifying and detecting high-risk transactions, which could benefit companies that wish to safeguard their reputation or to comply with local regulations.

Our findings spark a discussion on the societal implications of reducing Bitcoins anonymity. Privacy is a fundamental human right, integral to the functioning of democracy, as it limits power of the government and private sector over the public. At face value, our work seems to attack the privacy of Bitcoin. However, making known such non-trivial weakspots of Bitcoins anonymity, as found in this work, can have positive societal implications: they make users aware of the privacy weaknesses, enabling them to prevent unintended identity disclosure and/or surveillance, motivate stakeholders to improve Bitcoin's underlying technology to increase privacy and foster the research on cryptocurrency anonymity. Moreover, a more transparent Bitcoin Blockchain could heighten the mainstream's trust in the cryptocurrency by enabling law enforcement to more easily track down criminals and thus discourage their use of Bitcoin for illicit activity. As the EU points out in their proposed cryptocurrency regulation: *"The credibility of virtual currencies will not rise if they are used for criminal purposes. In this context, anonymity will become more a hindrance than an asset for virtual currencies taking up and their potential benefits to spread* [2].

7. Conclusion and Future Work

In this paper, a multi-class classification on Bitcoin Blockchain clusters is conducted. The aim was to investigate whether one can predict the category of a yet-unidentified cluster, given a set of already identified clusters serving as training data. The results show, that by utilising already identified, clustered and categorised addresses, it is possible to predict the type of a yet-unidentified cluster with an accuracy of 77% and an F1-score of 0.75 using the Gradient Boosting Classifier. The outcome of the research demonstrates, that the assumed level of anonymity of the Bitcoin Blockchain is not as high as commonly believed and the number of potential owners of a Bitcoin address can be narrowed down to a certain degree. This work paves the way for further research, where an increased amount of data and alternative classification approaches may lead to improved results.

In the future, we would seek to increase the relatively low sample size of identified clusters and add further cluster categories to create a more fine-grained differentiation between the clusters. Also, additional data could be utilised by harnessing more of the inherently available data on the Bitcoin Blockchain, as discussed in Sec. 4.2. Also, the feature engineering process could be improved, e.g. by using automated feature extraction. Lastly, we want to apply our model on the whole of Bitcoin Blockchain data and consequently present insights on the uncovered structure of the Bitcoin Blockchain, such as category distribution and transaction flow characteristics between those distributions.

[2] Council and Parliament of the European Union: Amendment to directive (eu) 2015/849

References

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

[2] S. T. Ali, D. Clarke, and P. McCorry, "Bitcoin: Perils of an unregulated global p2p currency," in *Cambridge International Workshop on Security Protocols*, Springer, 2015.

[3] H. Karlstrøm, "Do libertarians dream of electric coins? the material embeddedness of bitcoin," *Distinktion: Scandinavian Journal of Social Theory*, vol. 15, no. 1, pp. 23–36, 2014.

[4] R. Böhme, N. Christin, B. Edelman, and T. Moore, "Bitcoin: Economics, technology, and governance," *The Journal of Economic Perspectives*, vol. 29, no. 2, pp. 213–238, 2015.

[5] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, (New York, NY, USA), pp. 213–224, ACM, 2013.

[6] J. Martin, *Drugs on the dark net: How cryptomarkets are transforming the global trade in illicit drugs*. Springer, 2014.

[7] M. C. V. Hout and T. Bingham, "?silk road?, the virtual drug marketplace: A single case study of user experiences," *International Journal of Drug Policy*, vol. 24, no. 5, pp. 385 – 391, 2013.

[8] J. Martin, "Lost on the silk road: Online drug distribution and the ?cryptomarket?," *Criminology & Criminal Justice*, vol. 14, no. 3, pp. 351–367, 2014.

[9] 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," in *Proceedings of the 2013 conference on Internet measurement conference*, pp. 127–140, ACM, 2013.

[10] F. Reid and M. Harrigan, "An analysis of anonymity in the bitcoin system," in *Security and privacy in social networks*, pp. 197–223, Springer, 2013.

[11] Chainalysis, "Protecting the integrity of digital assets." https://www.chainalysis.com/, 2017.

[12] E. Androulaki, G. O. Karame, M. Roeschlin, T. Scherer, and S. Capkun, "Evaluating user privacy in bitcoin," in *International Conference on Financial Cryptography and Data Security*, pp. 34–51, Springer, 2013.

[13] D. Ron and A. Shamir, "Quantitative analysis of the full bitcoin transaction graph," in *International Conference on Financial Cryptography and Data Security*, pp. 6–24, Springer, 2013.

[14] P. Koshy, D. Koshy, and P. McDaniel, "An analysis of anonymity in bitcoin using p2p network traffic," in *International Conference on Financial Cryptography and Data Security*, pp. 469–485, Springer, 2014.

[15] M. Fleder, M. S. Kester, and S. Pillai, "Bitcoin transaction graph analysis," *arXiv preprint arXiv:1502.01657*, 2015.

[16] J. Hirshman, Y. Huang, and S. Macke, "Unsupervised approaches to detecting anomalous behavior in the bitcoin transaction network," 2013.

[17] C. Zhao, "Graph-based forensic investigation of bitcoin transactions," 2014.

[18] K. Liao, Z. Zhao, A. Doupé, and G.-J. Ahn, "Behind closed doors: measurement and analysis of cryptolocker ransoms in bitcoin," in *Electronic Crime Research (eCrime), 2016 APWG Symposium on*, pp. 1–13, IEEE, 2016.

[19] J. D. Nick, *Data-Driven De-Anonymization in Bitcoin*. PhD thesis, ETH-Zürich, 2015.

[20] S. Barber, X. Boyen, E. Shi, and E. Uzun, "Bitter to better-how to make bitcoin a better currency," in *Financial cryptography and data security*, pp. 399–414, Springer, 2012.

[21] J. Bonneau, A. Narayanan, A. Miller, J. Clark, J. A. Kroll, and E. W. Felten, "Mixcoin: Anonymity for bitcoin with accountable mixes," in *International Conference on Financial Cryptography and Data Security*, pp. 486–504, Springer, 2014.

[22] E. B. Sasson, A. Chiesa, C. Garman, M. Green, I. Miers, E. Tromer, and M. Virza, "Zerocash: Decentralized anonymous payments from bitcoin," in *Security and Privacy (SP), 2014 IEEE Symposium on*, IEEE, 2014.

[23] S. Meiklejohn and C. Orlandi, "Privacy-enhancing overlays in bitcoin," in *International Conference on Financial Cryptography and Data Security*, pp. 127–141, Springer, 2015.

[24] Bitcoin.org, "Protect your privacy." https://bitcoin.org/en/protect-your-privacy, 2009.

[25] M. Spagnuolo, F. Maggi, and S. Zanero, *BitIodine: Extracting Intelligence from the Bitcoin Network*, pp. 457–468. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168, ACM, 2006.

[29] P. Kennedy, *A guide to econometrics*. MIT press, 2003.

[30] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.

[31] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[32] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2012.